DOCUMENT RESUME

ED 464 934 TM 033 865

AUTHOR Pommerich, Mary

TITLE The Effect of Administration Mode on Test Performance and Score Precision, and Some Factors Contributing to Mode Differences.

PUB DATE 2002-04-00

NOTE 67p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 2-4, 2002).

PUB TYPE Numerical/Quantitative Data (110) -- Reports - Research (143) -- Speeches/Meeting Papers (150)

EDRS PRICE MF01/PC03 Plus Postage.

DESCRIPTORS Achievement Tests; Adaptive Testing; *Computer Assisted Testing; Computer Literacy; *High School Students; High Schools; Scores; *Test Format; Test Results; Testing Problems

IDENTIFIERS Item Parameters; *Paper and Pencil Tests

ABSTRACT

This paper considers differences in modes of test administration, addressing three questions: (1) Do examinees respond to items in the same way across administration modes and computer interface variations? (2) What are some of the factors that can contribute to modal effects? and (3) Can item parameters calibrated from paper and pencil administrations be used for computer administrations? The questions were examined using data from paper and pencil and computer administrations of a fixed-form test in two different examinee samples. Several of the tests studied were complex. An initial comparability study was performed in 1998 involving approximately 8,600 students, and in response to that study, revisions were made to the computer interfaces. A follow-up study was performed in 2000 with approximately 12,000 examinees. This paper examines performance differences across paper and computer modes and across computer interface variations in both studies. Results are summarized at the total test level and for some individual items. Some factors that might have contributed to mode differences or affected computer performance in general are discussed. A small simulation study was also performed to examine the effect of using item parameters calibrated from paper and pencil administrations in a computer administration. Some items showed no performance differences across administration modes, but other items did. A variety of factors appeared to contribute to mode effects, and each item seemed to have a unique set of circumstances. Changes to the computer interface appeared to affect performance. Overall, results suggest that while performance effects do occur across modes, they have a fairly small effect in practice. Simulation results suggest that item parameters calibrated from paper and pencil tests could probably be used initially in a computer administration. (Contains 18 tables.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 464 934

M. Pommerich

1

U.S. DEPARTMENT OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

The Effect of Administration Mode on Test Performance and Score Precision, and Some Factors Contributing to Mode Differences

Mary Pommerich
Defense Manpower Data Center

TM033865

Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA, April 2002.

BEST COPY AVAILABLE

Acknowledgements

Many people contributed to the planning, design, and data collection for the studies presented in this paper, including members of the Support, Technological Applications and Research Department, the Elementary and Secondary School Programs Department, the Measurement Research Department, the Operations Department, the Educational Services Department, and the Statistical Research Department, all at ACT.

Many people contributed to the design and development of the computer interfaces and tutorials used in the studies presented in this paper, including members of the Support, Technological Applications and Research Department, the Elementary and Secondary School Programs Department, the Placement Programs Department, the Systems Support Department, and the Educational Technology Center, all at ACT.

Many people contributed to data analyses that were conducted for the studies presented in this paper, including members of the Measurement Research Department and the Support, Technological Applications and Research Department, all at ACT.

In particular, thanks to Dean Colton and Han-wei Chen for identifying problematic records and cleaning the datasets discussed in this paper. Thanks to Jim Patterson, David Duer, Ann Gordon, and Beth Gehring for developing hypotheses presented in this paper. Thanks to Brad Hanson for supplying the results from the equating study presented in the paper. Finally, thanks to Dan Segall for his assistance on the simulation portion of the paper.

The Effect of Administration Mode on Test Performance and Score Precision, and Some Factors Contributing to Mode Differences

As testing moves from paper and pencil administration toward computerized administration, how to present complex tests on a computer screen becomes an important concern. Information that can be viewed in full on a two-page spread in a booklet cannot typically be presented on a single computer screen. In a dual-platform testing program with a complex test, taking certain items in one mode or the other could possibly advantage some examinees. Even in a computer-only platform, decisions about how to present the test could affect examinee performance. Seemingly subtle differences in how the test is presented on computer could have a not-so-subtle effect on examinee performance.

Computerized administration is less of an issue for discrete-item tests such as Math, if single items can be presented in full on a computer screen. Computerized administration is more of an issue for complex tests that contain information that cannot all be displayed on-screen at once for an item. For example, a test with long text-based passages is complex if the examinee must navigate through the passages to read and find answers to items. A test with text-based passages containing multiple figures or tables per passage is complex, particularly if the figures and tables need to be compared. As computerized presentation of tests becomes more of a reality, it is important to develop an understanding of the presentation choices we make and how they can affect an examinee's performance. Presenting a complex test on computer is not an easy task; many decisions need to be made about how best to present the information, so that the method of presentation does not interfere with examinee performance on the test.

Because of potential mode effects, Parshall, Spray, Kalohn, and Davey (2002) suggest that testing programs that treat scores across different administration platforms as equivalent should perform studies to document the comparability of the test scores. Mode effects also are an important consideration if items are calibrated in one medium, and then used operationally in another medium. For example, when starting a computerized testing program, it may be very costly and time-consuming to calibrate the initial pool(s) using data from computer administrations of the items. Every item in the pool would have to be administered to a sufficient number of examinees in order to calibrate. If the item pool is large, this would require a substantial amount of testing that likely cannot be done quickly (or cheaply) via computer administration. Thus, a testing program might consider initially using item parameters calibrated

from paper and pencil administrations of the items for operational computer administrations, until enough data are obtained to calibrate from the computer administrations. Parshall, et al. (2002) caution that item calibrations based on paper and pencil administrations might not represent the performance of those same items in a computer administration.

This paper addresses three questions:

(1) Do examinees respond to items in the same way across administration modes and computer interface variations?

(2) What are some of the factors that can contribute to mode effects?

(3) Can item parameters calibrated from paper and pencil administrations be used for computer administrations?

The questions are examined using data from paper and pencil and computer administrations of a fixed-form test in two different examinee samples. Several of the tests studied were complex. An initial comparability study was performed in 1998. In response to findings from that study, revisions were made to the computer interfaces, and then a follow-up comparability study was performed in 2000. The paper examines performance differences across paper and computer modes and across computer interface variations in both studies. Results are summarized at the total test level and for some individual items. Some factors that might have contributed to mode differences or affected computer performance in general are discussed. In addition, a small simulation study was performed to examine the effect of using item parameters calibrated from paper and pencil administrations in a computer administration.

Description of the Tests, Computer Interfaces, and Comparability Studies

Two comparability studies were performed in 1998 and 2000, called Comparability 1 and Comparability 2, respectively. In each study, the same fixed-form tests were administered across paper and pencil and computer modes in the content areas of English, Math, Reading, and Science Reasoning. Slightly different Math tests were used across the two comparability studies. As such, results from Math are not presented here. An initial computer interface was used in Comparability 1 and then modified for Comparability 2 based on findings from Comparability 1. The interface used in Comparability 1 is referred to as Interface 1. The interface used in Comparability 2 is referred to as Interface 2.

English

Test Content

The English test consisted of four passages containing underlined words and phrases, with 15 multiple-choice items in each passage (60 items total). For most items, examinees were instructed to choose the response option for the underlined portion that best expressed the idea, made the statement appropriate for standard written English, or was worded most consistently with the style and tone of the passage as a whole. These types of items had no stimulus associated with them (i.e., there were only response options, and no preceding question). For some items, there was a stimulus present that asked a question about the underlined portion in the passage. Examinees were instructed to choose the best answer to the question.

Booklet Presentation

In the booklet presentation of the English test, the passage and items were presented jointly on a page. The passage was presented in the left half of the page, while the items were presented in the right half of the page. Each underlined portion was always aligned with the top of the item. The passages and accompanying items occupied about two booklet pages each. Examinees were able to move freely throughout all English passages and items in the booklet while taking the English test. They could respond to items and passages in any order, and were not required to give responses to all items. Similar rules of movement between items and passages held for the Reading, and Science Reasoning paper and pencil tests. Within a single test, examinees were allowed to move freely throughout the test.

Computer Presentation

In the computer presentation for both Interface 1 and Interface 2, the passage and items were presented jointly on the screen, with the passage on the left half and the items on the right half of the screen. The passage was not visible in its entirety on the computer screen. The examinee had to scroll through the passage to see the passage in its entirety, although the passage automatically scrolled for examinees on various items (see further discussion below). Items were presented one at a time. Within a passage, examinees were allowed to answer items in any order. They were required to answer all items prior to moving on to the next passage. Once an examinee completed a passage and moved on to the next passage, they were not allowed to return to the previous passage. Also, passages were presented one at a time, so that examinees could not see the next passage until they proceeded to it. A similar presentation of the passage

and item windows was used with the computerized Reading and Science Reasoning tests, along with the same rules for moving between items and passages.

Computer Interface Features

In Interface 1, the following features were utilized:

- The full underlined portion in the passage window was highlighted.
- The passage automatically scrolled when an item was selected that was not visible on screen (about every 6th item).
- The underlined portions were not aligned with the top of the question.

Results from Comparability 1 (to be discussed in more detail later) showed that whereas some individual items favored computer examinees and some individual items favored paper examinees, as a whole, the test tended to favor computer examinees. After a review of the test content, test booklet, computer interface, and discussions with examinees, the following hypotheses were posited as possible explanations for why computer examinees performed better overall than paper examinees (see Pommerich & Burden (2000) for further discussion):

- The use of full highlighting was advantageous to computer examinees as it drew their attention to the full underlined portion.
- Computer examinees on some items were better able to focus on relevant sections of passages/items because those sections were centered in the passage and item windows and examinees were not distracted by extraneous information presented in the rest of the test. This phenomenon will be referred to as the "focus effect."

Results also suggested the following hypotheses:

- Computer examinees might have been less likely to read the stimulus preceding the response options, for items containing a stimulus.
- Where the underlined portion was aligned with the response options might influence the response selected.

Thus, the following changes were implemented for Interface 2:

- The full highlighting of the underlined portion was removed. Instead, only the item number underneath the underlined portion was highlighted.
- The item number was placed adjacent to the top of the question within the item window to match what is done in the booklets (in Interface 1, the items were not numbered adjacent to the question.)

- Two automatic scrolling variations were compared:
    - The passage scrolled when an item was selected that was not visible on screen (about every 6th item), so that the underlined portion was not always aligned with the top of the question. This scrolling was used in both Interface 1 and Interface 2. This condition will be referred to as *English Semi*.
    - The passage scrolled every time a new item was selected, so that the underlined portion was always aligned with the top of the question. This scrolling was only used in Interface 2. This condition will be referred to as *English Auto*.

Reading

Test Content

The Reading test consisted of four passages with 10 multiple-choice items on each passage (40 items total). Examinees were instructed to read the passage and choose the best answer to each question. Items on the Reading test generally fell into two types: questions that required a global understanding of the passage and questions that required knowledge of specific information given in the passage. For global questions, examinees typically had to make an inference from what they had read to answer the question. Some of the items had line references associated with them (i.e., the item stimulus contained the number of a line or lines in the passage to which they were directed to read). In the booklet presentation, the reading passage was presented first in its entirety, in two columns per page. The passages were followed by the test items. The passages and accompanying items occupied about two booklet pages each. The computer presentation for Reading corresponded to that described for the English test.

Computer Interface Features

In Interface 1, the following features were utilized:

- Examinees moved through the passage by scrolling.
- Examinees could scroll line-by-line, or use a sliding scroll bar to move quickly through the passage.
- Line-by-line scrolling speed was not very fast.
- Pre-test training for scrolling options was for line-by-line scrolling only (examinees were not explicitly shown how to use the sliding scroll bar).
- Line breaks were not the same as in the booklet, so the content of referenced lines was not the same across modes.

Results from Comparability 1 (to be discussed in more detail later) showed that whereas some individual items favored computer examinees and some individual items favored paper examinees, as a whole, the test tended to favor paper examinees. After a review of the test content, test booklet, computer interface, and interviews with examinees, the following hypotheses were posited as possible explanations for why paper examinees performed better overall than computer examinees (see Pommerich & Burden (2000) for further discussion):

- Computer examinees sometimes had difficulty locating information in the passage with scrolling as the navigation method.
- Paper examinees might have been more likely than computer examinees to experience "positional memory," whereby they remembered the location of information given in the passage, because the passage occurred in a fixed position on the page.
- Slow scrolling speed was a hindrance for computer examinees.
- Different line breaks could have created mode differences on questions with line references.

Thus, the following changes were implemented for the Interface 2:

- Line breaks for the passages were made the same across booklet and computer representations, so that each line contained the same content across modes.
- Scrolling speed was increased.
- Examinees were explicitly taught to use the sliding scroll bar prior to testing.
- Two navigation variations were compared:
    - Examinees moved through the passage by scrolling, using either line-by-line scrolling or a sliding scroll bar. This scrolling was used in Interface 1 and Interface 2 (although scrolling speed was increased and pre-test instruction on scrolling was more comprehensive for Interface 2). This condition will be referred to as *Read Scroll*.
    - Examinees moved through the passage by paging. In this variation, the passage was divided into separate pages and the examinee moved between pages by clicking on a specific page number, or by using "Next Page" or "Previous Page" buttons. Paging was only used in Interface 2. This condition will be referred to as *Read Page*.

Science Reasoning

Test Content

- The Science Reasoning test consisted of seven passages with varying numbers of multiple-choice items per passage (5-7 items per passage; 40 items total). Some passages contained figures and tables. In the booklet presentation, the passage was presented first in its entirety, in two columns per page. The passages and accompanying items occupied about two booklet pages each. The passages were followed by the test items. The computer presentation for Science Reasoning corresponded to that described for the English test, with the additional feature that some figures and tables within the passage were enlargeable and moveable.

Computer Interface Features

In Interface 1, the following features were utilized:

- Examinees moved through the passage by scrolling.
- Examinees could scroll line-by-line, or use a sliding scroll bar to move quickly through the passage.
- Line-by-line scrolling speed was not very fast.
- Pre-test training for scrolling options was for line-by-line scrolling only (examinees were not explicitly shown how to use the sliding scroll bar).

Results from Comparability 1 (to be discussed in more detail later) showed some individual items favoring computer examinees and some individual items favoring paper examinees. Overall, there was no trend in results, although the last passage (Passage 7) favored paper examinees, and Passage 4 favored computer examinees. After a review of the test content, test booklet, computer interface, and interviews with examinees, the following hypotheses were posited as possible explanations for why computer and paper examinees performed differently on individual items/passages (see Pommerich & Burden (2000) for further discussion):

- Computer examinees sometimes had difficulty locating information given in the passage with scrolling as the navigation method.
- Paper examinees might have been more likely than computer examinees to experience "positional memory," whereby they remembered the location of information in the passage, because the passage occurred in a fixed position on the page.
- Slow scrolling speed was a hindrance for computer examinees.

- Computer examinees had difficulty comparing information across two tables or figures, and were unaware that they could move an enlarged graphic so that it could be viewed at the same time as another graphic.
- Computer examinees were advantaged by a "focus effect" on some items (i.e., they were better able to focus on relevant sections of passages/items because those sections were centered in the passage and item windows and examinees were not distracted by extraneous information presented in the rest of the test).

Thus, the following changes were implemented for Interface 2:

- Scrolling speed was increased.
- Examinees were explicitly taught to use the sliding scroll bar prior to testing.
- Two graphics were allowed to be enlarged simultaneously and moved so they could be viewed side-by-side.
- Two navigation variations were compared:
    - Examinees moved through the passage by scrolling, using either a line-by-line scrolling or a sliding scroll bar. This scrolling was used in Interface 1 and Interface 2 (although scrolling speed was increased and pre-test instruction on scrolling was more comprehensive in Interface 2). This condition will be referred to as *Science Scroll*.
    - Examinees moved through the passage by paging. In this variation, the passage was divided into separate pages and the examinee moved between pages by clicking on a specific page number, or by using "Next Page" or "Previous Page" buttons. Paging was only used in Interface 2. This condition will be referred to as *Science Page*.

Changes Between Interface 1 and Interface 2 for All Tests

The following changes were implemented between Interface 1 and Interface 2 the same way over all tests unless otherwise indicated:

- Changed the wording on some buttons and on text adjacent to the buttons to be more concise and clear.
- Different colors and button designs were used to change the look of the interface.
- Additional passage and item numbering was added to the outside of the passage and item windows, to clarify which item/passage the examinee was on (i.e., indicated Passage 1 of 4,

Question 1 of 60, etc. In Interface 1, the current passage and question number was given, but no information was given as to how many passages or questions remained.)

- On startup of a passage, the first question was not displayed until the examinee selected the first item, to encourage examinees to read the passage before answering the first question.

The Comparability Studies

Comparability 1

Comparability 1 compared performance across computer and paper and pencil administrations of the same fixed form, using computer Interface 1. Testing was conducted between September and December in 1998. A total of 40 schools participated in the study, with approximately 8600 students testing overall. Within a school, examinees were randomly assigned to a paper and pencil or computer administration of a fixed-form test. Within each administration mode, examinees were randomly assigned to one of the following content areas: English, Math, Reading, or Science Reasoning. (Note that only one computer interface variation was used in each content area.) Thus, there were a total of eight administration conditions. All computer examinees took a short tutorial prior to testing that demonstrated how to use all of the functions necessary to take the computerized test (with the exception of demonstrating the use of the sliding scroll bar, as discussed earlier). The fixed forms were drawn from an intact paper and pencil form. Reading and Science Reasoning were administered in their entirety with the same time constraints as used operationally, while a representative subset of items was selected from the English and Math tests to accommodate a 35-minute testing period. Total testing time was 35 minutes for all content areas and modes.

Comparability 2

Comparability 2 compared performance across computer and paper and pencil administrations of the same fixed form, using computer Interface 2. Interface 2 was a modified version of Interface 1, developed in response to findings from Comparability 1. Testing was conducted between October 2000 and January 2001. A total of 61 schools participated in the study, with approximately 12,000 examinees testing. Within a school, examinees were randomly assigned to a paper and pencil or computer administration of a fixed-form test. Examinees assigned to the paper mode were randomly assigned to one of the following content areas: English, Math, Reading, or Science Reasoning. Examinees assigned to the computer mode were randomly assigned to one of the following content area and interface variations: English Auto,

English Semi, Math, Reading Scroll, Reading Page, Science Reasoning Scroll, or Science Reasoning Page. Thus, there were a total of 11 administration conditions. All computer examinees took a short tutorial prior to testing that demonstrated how to use all of the functions necessary to take the computerized test. The tutorial used in Comparability 2 was more comprehensive and more interactive than the tutorial used in Comparability 1.

The same fixed form test was administered across computer and paper and pencil administration modes. The fixed forms were originally drawn for Comparability 1 from an intact paper and pencil form. Reading and Science Reasoning were administered in their entirety with the same time constraints as used operationally, while a representative subset of items was selected for English to accommodate a 35-minute testing period. The Math test was further modified to include some special item types not used when the form was first administered operationally. Total test time for all content areas and modes was 35 minutes. The test forms for English, Reading, and Science Reasoning were identical to those used in Comparability 1.

Results

Data Cleaning

Due to irregularities during assignment to a testing condition or during testing itself, some records were unusable; records that were problematic were deleted from the final analyses.[1] The final sample sizes for the analyses are reported in Table 1.[2] Within Comparability 1 and Comparability 2, these groups are considered to be randomly equivalent.

Completion Rates

A concern in computerizing a paper and pencil test is that it might take more time for examinees to complete the test on computer than on paper. If it takes computer examinees longer to complete the same test than paper examinees, that is potentially unfair to computer examinees. Many factors could contribute to an increased testing time for computer examinees. It may take more time to use a mouse to navigate and respond to questions. It may also take more time simply to find information on a computer since it is not all visible on screen at once. It may take more time to navigate to find information on the computer. It may be more difficult to read from a computer screen than a test booklet. If it takes more time to take the same test on computer than on paper, then longer testing times may need to be allocated for testing on

[1] Many more irregularities were observed in Comparability 1 than in Comparability 2.
[2] Results for Math will not be discussed.

computer than testing on paper. It would be advantageous to design a computer interface to minimize the time needed to yield comparable scores across modes, since computer examinees would also have to take additional time to complete a tutorial. The percentages of examinees finishing the test are given in Table 2, for both Comparability 1 and Comparability 2.

Table 1. Final Sample Sizes for Analyses

	Comparability 1		Comparability 2	
Test	Condition	N	Condition	N
English			Computer Auto	1110
	Computer	905	Computer Semi	1031
	Paper	1040	Paper	1137
Math	Computer	918	Computer	1083
	Paper	994	Paper	1099
Reading			Computer Page	996
	Computer	908	Computer Scroll	1089
	Paper	985	Paper	1086
Science Reasoning			Computer Page	902
	Computer	827	Computer Scroll	1067
	Paper	947	Paper	1055

For all content areas, the completion rates for the paper mode were lower in Comparability 2 than in Comparability 1. If there were no sample differences, we would expect completion rates to be about the same across studies because the same forms and testing time were used. The lower completion rates in Comparability 2 are likely a result of having a less academically able sample than in Comparability 1, arising from a greater solicitation of less academically able schools for Comparability 2.[3] We would expect then, that if our interface changes did not have any effect on reducing the time needed to complete the test, completion rates for computer examinees in Comparability 2 would be lower than completion rates for computer examinees in Comparability 1 simply because of the sample differences across the two studies. Because the computer completion rates are about the same or higher for Comparability 2 than Comparability 1, this suggests that the interface changes in general decreased the amount of time needed to complete the test on computer. In addition, completion rates for paper examinees might have been inflated somewhat if examinees answered quickly and randomly at

[1] For examinees matched to an ACT Assessment score, the average ACT Reading and Science Reasoning scores were at least one scale score point higher for Comparability 1 examinees than for Comparability 2 examinees.

the end just to have a response to all items. It is likely easier for paper examinees to complete the test in such a way than computer examinees. As such, completion rates for paper examinees might appear higher than completion rates for computer examinees.

Table 2. Percent Completing the Test Across Comparability 1 and Comparability 2

	Comparability 1		Comparability 2	
Test	Condition	Percent Completing Test	Condition	Percent Completing Test
English	-		Computer Auto	83.42
	Computer	81.00	Computer Semi	81.77
	Paper	81.00	Paper	78.36
Reading	-		Computer Page	64.16
	Computer	64.20	Computer Scroll	62.72
	Paper	76.20	Paper	70.99
Science	-		Computer Page	65.19
Reasoning	Computer	56.00	Computer Scroll	63.64
	Paper	68.80	Paper	60.86

English

Completion rates were the same for paper and computer examinees in Comparability 1. Completion rates for both computer conditions in Comparability 2 were higher than the paper completion rates. Completion rates for the Auto condition were slightly higher than the completion rates for the Semi condition. There was a lot of white space in the English passage between adjacent lines, and examinees generally did not have to scroll while responding to individual items. As a result, it might have been somewhat easier and quicker to focus on information on the computer than on paper if it was contained within the screen, and extraneous information was hidden from view. In all, computer results appeared less speeded for Comparability 2 than for Comparability 1. It is likely that examinees in Comparability 2 had a greater awareness of where they were in the test and how much time remained than examinees in Comparability 1, due to the inclusion of more information to that effect in the revised interface. It is also likely that navigation was improved through better training on how to navigate in the tutorial.

Reading

Completion rates were much higher for paper than for computer in Comparability 1. Completion rates were still higher for paper than for both computer conditions in Comparability

2, although the difference in completion rates was smaller, particularly for the Page condition. Taking into consideration sample differences across the two studies, it appears that the Reading computer results may have been less speeded in Comparability 2 than they were in Comparability 1. It is likely that examinees in Comparability 2 had a greater awareness of where they were in the test and how much time remained than examinees in Comparability 1, due to the inclusion of more information to that effect in the revised interface. It is also likely that navigation was improved through increased scrolling speed (in the Scroll condition), the use of paging (in the Page condition), and better training on how to navigate in the tutorial. The higher completion rates for paper in Comparability 2 likely still occurred because the Reading passages were very dense. There was a lot of text that examinees had to wade through to find information, which could have been difficult to do on computer. It might be difficult to match the paper completion rates on the computer without making the passages shorter, increasing testing time, or creating more white space between the lines of the passage.

Science Reasoning

Completion rates were much higher for paper than for computer in Comparability 1. In Comparability 2, the results flip-flopped and the completion rates were higher for both computer conditions than for the paper condition. Completion rates were slightly higher for the paging condition than the scrolling condition. Higher completion rates on the computer for Comparability 2 examinees over Comparability 1 examinees could again be explained by Comparability 2 examinees having a greater awareness of where they were in the test and how much time remained than Comparability 1 examinees. It is also likely that navigation was improved through increased scrolling speed (in the Scroll condition), the use of paging (in the Page condition), and better training on how to navigate in the tutorial. Improved navigation is also a plausible explanation for the higher completion rates for computer examinees over paper examinees in Comparability 2. Science Reasoning examinees sometimes had to compare information across two figures on tables that could not be viewed simultaneously on the computer screen. Comparability 1 examinees likely were severely hampered by the slow navigation capabilities of Interface 1 when moving back and forth between figures and tables. The improved navigation of Interface 2 appears to have had a substantial effect on completion rates for Comparability 2 computer examinees.

The higher completion rates for Comparability 2 computer examinees relative to paper examinees might also be attributable in part to the "focus effect" discussed earlier. It might be easier to focus on information on the computer than on paper if it is all contained within the screen, and extraneous information is hidden from view. The fact that the results are the opposite as those observed for Reading (i.e., completion rates for Reading computer examinees in Comparability 2 are still below the completion rates of the paper examinees) suggests that even with improved navigation and training, there may be no focus effect for Reading computer examinees. This might be due to the nature of the information contained in the passage and how it is presented. While the passages are lengthy in Science Reasoning, the sheer density of text compared to the Reading passages is much less because the inclusion of figures and tables creates more white space surrounding text in the passage. Even with improved navigation, it might be more difficult to find where information is located in Reading than in Science Reasoning because of the density of text contained in the Reading passage.

Total Score Performance

Average number right scores for each test and condition are given in Table 3, for both Comparability 1 and Comparability 2. The average scores were lower in Comparability 2 than in Comparability 1, as might be expected if the Comparability 2 examinees were less academically able. Table 4 gives the difference in average number right scores across modes for each computer condition (computer – paper). A positive difference indicates a higher score on computer than on paper.

Table 3. Average Number Right Scores Across Comparability 1 and Comparability 2

	Comparability 1		Comparability 2	
Test	Condition	Average Score	Condition	Average Score
English (60 Items)	-		Computer Auto	34.03
	Computer	36.09	Computer Semi	33.80
	Paper	34.90	Paper	32.38
Reading (40 Items)	-		Computer Page	20.16
	Computer	21.08	Computer Scroll	20.12
	Paper	22.13	Paper	20.37
Science Reasoning (40 Items)	-		Computer Page	21.97
	Computer	23.01	Computer Scroll	21.68
	Paper	23.07	Paper	21.24

Table 4. Difference in Average Number Right Score Across Modes (Computer – Paper)

	Comparability 1		Comparability 2	
Test	Condition	Difference	Condition	Difference
English	-		Auto	+1.65
	Computer	+1.19	Semi	+1.42
Reading	-		Page	-0.21
	Computer	-1.05	Scroll	-0.25
Science	-		Page	+0.73
Reasoning	Computer	-0.06	Scroll	+0.44

For English, computer examinees scored higher on average than paper examinees under both computer conditions in Comparability 2. Of the two computer conditions, Auto examinees scored slightly higher than Semi examinees. Computer examinees also scored higher than paper examinees in Comparability 1. The difference in average scores across modes was larger for Comparability 2 than Comparability 1, so there was a widening of the performance gap favoring computer examinees across the two studies.

For Reading, computer examinees scored lower on average than paper examinees for both computer conditions in Comparability 2. Of the two computer conditions, Page examinees scored slightly higher than Scroll examinees. Computer examinees also scored lower than paper examinees in Comparability 1. The difference in average scores was much smaller for Comparability 2 than Comparability 1, so there was a narrowing of the performance gap favoring paper examinees across the two studies.

For Science Reasoning, computer examinees scored higher on average than paper examinees for both computer conditions in Comparability 2. Of the two computer conditions, Page examinees scored slightly higher than Scroll examinees. In Comparability 1, computer examinees scored slightly lower than paper examinees. The difference in average scores was larger for Comparability 2 than Comparability 1, with a shift in direction from slightly favoring paper examinees to favoring computer examinees. This trend complements the shift in completion rates for Science Reasoning noted earlier.

For all content areas, it is likely that improvements in Interface 2 and the tutorial made it easier for examinees to use the interface, navigate throughout the test, and respond more quickly, leading to an improved performance of computer examinees relative to paper examinees in Comparability 2 over Comparability 1.

Item Level Performance

English

Plots of individual item p-value differences with error bands (computer p-value – paper p-value ± 2 standard errors) across paper and computer conditions are given in Figures 1-2. Each passage is separated by a vertical line in the plots. Figure 1 shows the computer – paper p-value differences for the English Auto and English Semi computer conditions from Comparability 2. Figure 2 shows the computer – paper p-value differences for Comparability 1 and for a baseline comparison based on two mutually exclusive random samples of examinees who took the items used in the comparability studies on paper as part of an equating study. (The items used for English in Comparability 1 and Comparability 2 were a subset of items from an intact form previously administered in an equating study.) The two groups in the baseline comparison are considered to be randomly equivalent, so one group was arbitrarily assigned to represent the "computer" condition, while the other was assigned to represent the "paper" condition. The sample sizes for the computer and paper sample were fixed at those observed for the respective condition in Comparability 1. Since these two groups took the same form via the same administration mode and were randomly equivalent, any performance differences are likely due to sampling error. By chance, we would expect a small percentage of items from the baseline comparison to display a significant performance difference.

In each of these plots, a positive difference indicates the item was easier on computer than on paper. We would expect that if there was no difference in performance across modes the error bands would surround zero (i.e., zero would not fall outside of the span of the error bars). Table 5 gives the number (and percent) of items for which the error bands do not surround zero in Comparability 1 and Comparability 2, for all content areas.

Results for Comparability 1 show four English items (7%) favoring paper examinees, and 16 items (27%) favoring computer examinees. Results for the Auto condition of Comparability 2 show 23 items (38%) favoring computer examinees and one item (2%) favoring paper examinees. Results for the Semi condition of Comparability 2 showed 21 items (35%) favoring computer examinees and one item (2%) favoring paper examinees. For the English baseline comparison, four items (7%) showed a significant performance difference across groups. These are likely due to chance.

Table 5. Number (and Percent) of Items With Error Bands That Do Not Surround Zero.

	Comparability 1			Comparability 2		
Test	Condition	# Favoring Computer	# Favoring Paper	Condition	# Favoring Computer	# Favoring Paper
English	-			Auto vs. Paper	23 (38%)	1 (2%)
	Computer vs. Paper	16 (27%)	4 (7%)	Semi vs. Paper	21 (35%)	1 (2%)
Reading	-			Page vs. Paper	4 (10%)	9 (23%)
	Computer vs. Paper	0 (0%)	12 (30%)	Scroll vs. Paper	3 (8%)	8 (20%)
Science	-			Page vs. Paper	13 (33%)	5 (13%)
Reasoning	Computer vs. Paper	6 (15%)	7 (18%)	Scroll vs. Paper	10 (25%)	5 (13%)

Figure 1 shows a marked trend favoring computer examinees toward the end of the test for both the Auto and Semi conditions in Comparability 2. Of the last 25 items, 16 favored computer examinees in the Semi condition, and 18 favored computer examinees in the Auto condition. In Comparability 1, only eight of the last 25 items favored computer examinees. In addition, only one item out of the entire test favored paper examinees for both the Auto and Semi conditions. For the first 35 items of the test, five favored computer and one favored paper, for both the Auto and Semi conditions. For Comparability 1 (which used the same scrolling as the Semi condition, albeit slower), eight of the first 35 items favored computer, and four favored paper.

With respect to Interface 2 relative to Interface 1, the results seem to indicate fewer items favoring paper examinees overall, fewer items favoring computer examinees in the beginning of the test, and more items favoring computer examinees at the end of the test. The interface changes and improved tutorial might have created more parity across paper and computer administrations early in the test, before speeded response behavior kicked in. English had the highest completion rates of all content areas in both Comparability 1 and Comparability 2, but the average amount of time spent on each item suggests that there was some speeded response behavior later in the test. On average, Semi examinees spent 84.4 seconds per item on the first 35 items (83.2 seconds for Auto). For the last 25 items, however, Semi examinees spent, on average, only 25.8 seconds per item (25.6 seconds for Auto). Once examinees start to rush to complete the exam, it might be advantageous to take the test on computer rather than on paper, because of the ease in responding and moving quickly through items, and a greater ability to focus on the question at hand without being distracted by extraneous information. In discussions with examinees that expressed a preference testing on computer, many mentioned that they

preferred testing on computer because it was easier not having to bubble in the answers. The hypothesized ease of engaging in speeded response behavior on the computer relative to paper will subsequently be referred to as the "no-bubble effect."

One might ask why all items at the end of the test did not favor computer examinees for the Auto and Semi conditions? A more detailed evaluation of item and interface design features suggests that there were many different factors that contributed to performance differences across modes. Figure 3 shows p-value difference plots for English items 6, 10, 13, 17, 18, 22, and 30. Each item plot contains the computer – paper p-value differences ± 2 standard errors for Comparability 1, for the Semi and Auto conditions from Comparability 2, and for the baseline comparison. Hypotheses explaining performance differences observed in Comparability 1 were developed for these items by test specialists after a review of the test content, test booklet, interface features, and interviews with examinees (see Pommerich & Burden, 2000). Findings across the two studies will be discussed here, relative to the original hypotheses and the interface changes. Some possible explanations for the results are offered as speculations. These are only speculations, because it cannot be known with any certainty from these studies exactly what caused the performance differences.

Item 6. In the test booklet, a page break occurred in the middle of the sentence that contained the underlined portion for this item. Results for Comparability 1 significantly favored computer examinees. It was hypothesized that paper and pencil examinees might be inclined to incorrectly respond "No Change," which was Option A, if they didn't consider the full sentence in answering, while the full highlighting of the underlined portion in Interface 1 might have helped some computer examinees realize that "No Change" was not a valid option. (Approximately 19.4% of Comparability 1 paper examinees selected Option A versus 12.8% of Comparability 1 computer examinees.) In Interface 2, results for the Semi condition favored computer examinees, although the difference was not significant, while results for the Auto condition showed a smaller difference. (Approximately 20.6% of Comparability 2 paper examinees selected Option A, versus 18.5% of Comparability 2 Semi examinees and 19.2% of Comparability 2 Auto examinees.) More computer examinees selected Option A in Interface 2 than in Interface 1, for both the Auto and Semi conditions. The removal of the full highlighting, along with the presentation of an item number adjacent to Option A may have made Semi examinees more inclined to select Option A in Interface 2. Auto examinees performed similarly

to paper examinees. They may have been further inclined to select Option A than Semi examinees because of the alignment of the underlined portion with Option A (recall that the alignment in the Interface 2 Semi condition was the same as that used in Interface 1, whereas the alignment in the Interface 2 Auto condition matched that of the paper presentation).

Item 10. This was the first item on the test that contained a stimulus. Examinees needed to read the stimulus to understand how to respond to the question. Results for Comparability 1 significantly favored paper examinees. It was hypothesized that computer examinees were less likely to read the stimulus than paper examinees because the underlined portion was not aligned with the top of the item (i.e., with the stimulus), and there was no numbering of the item to draw their attention to the stimulus. In Interface 2, item numbering was added adjacent to the top of the item for both the Auto and Semi conditions. Results for Comparability 2 still significantly favored paper examinees for the Semi condition, but to a lesser degree than in Comparability 1. Thus, it appears that the inclusion of the item number may have helped draw the Semi examinees' attention to the stimulus, but that the lack of alignment of the underlined portion with the stimulus could still have caused some computer examinees to ignore the stimulus while responding. Results for the Auto condition, which aligned the underlined portion with the stimulus, also favored paper examinees, but the difference was not significant.

Item 13. All response options for this item looked acceptable. Results for Comparability 1 significantly favored computer examinees. It was hypothesized that the alignment of response options with the underlined portion may have influenced the computer examinees' responses. The correct answer was D, and the underlined portion was aligned below option D, so computer examinees may have been inclined to read the response options from bottom to top and select the first acceptable-looking response. In Interface 2, the underlined portion was also aligned below option D under the Semi condition. For the Auto condition, the underlined portion was aligned with the top of the response options. The results for the Semi condition in Comparability 2 significantly favored the computer examinees. The results for the Auto condition also favored the computer examinees, but the difference was not significant. It is possible that when the underlined portion falls below the response options, the examinees might be more inclined to read the options from bottom to top. When the underlined portion is aligned with the top of the item, the examinees might be more inclined to read the options from top to bottom.

Item 17. The underlined portion for this item contained a word that might have been unfamiliar to many examinees. Results for Comparability 1 significantly favored computer examinees. It was hypothesized that full highlighting of the underlined portion made the examinees focus on the underlined portion as a viable response option and choose "No Change," which happened to be the correct response. In Interface 2, the full highlighting was removed and only the item number underneath the underlined portion was highlighted. The results for both the Auto and Semi condition showed no favoritism for computer examinees under the new interface.

Item 18. This item contained a stimulus that examinees had to read to answer. It was difficult to guess correctly if the stimulus was not read. Results for Comparability 1 significantly favored paper examinees. It was hypothesized that computer examinees were less likely to read the stimulus because the underlined portion fell below the stimulus and there was a lack of focus on the stimulus. In Interface 2, it was hoped that the item number adjacent to the top of the question within the item window would help draw examinees' attention to the stimulus. For the Semi condition, the underlined portion was not aligned with the top of the stimulus, whereas for the Auto condition, the underlined portion was aligned with the top of the stimulus. The results for the Semi condition favored paper examinees, but not significantly. The results for the Auto condition didn't favor either group. Results for this item were fairly similar to results for Item 10, which also contained a stimulus. As in Item 10, where the underlined portion was aligned with the stimulus appeared to have a bigger influence on computer examinees' responses than the inclusion of the item number, although the item numbering may have helped somewhat.

Item 22. This item also contained a stimulus that examinees had to read to answer. The stimulus referred to the previous two sentences of the passage and asked examinees to select the response option that paralleled the style used in those sentences. Results for Comparability 1 significantly favored paper examinees. It was hypothesized that computer examinees were less likely to read the stimulus because of the alignment of the underlined portion with the stimulus. Unlike Item 18, results for Comparability 2 significantly favored Auto computer examinees, while the performance difference was not significant for the Semi condition. In Interface 1, it may have been difficult to infer the correct response without reading the stimulus, because of the layout of the relevant sentences on the screen. The pattern to follow was not apparent without carefully reading the two sentences because the pattern did not stand out given the layout of the

relevant section. In Interface 2 under the Semi condition, the underlined portion was still not aligned with the stimulus, but different line breaks were used from Interface 1, so the layout of the relevant section was different from Interface 1. Under this layout, the pattern used in the two sentences was more obvious, and it was likelier easier to imitate without reading the stimulus. Interface 2 Semi examinees might not have been any more likely to read the stimulus than Interface 1 examinees, but it could have been easier for them to infer the correct answer without reading the stimulus because of the layout of the relevant sentences.

Results significantly favoring computer Auto examinees could also be explained by the layout of the relevant sentences. Auto examinees had to scroll up to see both of the referenced sentences, and so may have been more likely to read the referenced sentences than paper examinees, if they read the stimulus. The results suggest that the alignment of the underlined portion with the stimulus in the Auto condition may have influenced computer examinees to read the stimulus, and thereby to scroll and read the referenced sentences.

Item 30. This item required the examinee to choose the correct tense for an underlined word. Results for Comparability 1 significantly favored computer examinees. It was hypothesized that the page layout in the booklet versus the computer layout affected performance. In the booklet, only the paragraph containing the underlined portion was visible on the page (which was the last paragraph in the passage), whereas on the computer the last two paragraphs of the passage were visible on screen when this item was selected. It is likely that with more of the passage visible, it was easier for computer examinees to correctly infer the tense of the passage. Results for Comparability 2 significantly favored computer examinees for Auto (although to a lesser degree than computer examinees were favored in Comparability 1), but not the Semi condition. In the Semi condition, only the last paragraph was visible on screen when this item was selected, which matched the layout in the test booklet. (Although the Semi condition scrolled at the same time as Interface 1 in Comparability 1, the line breaks differed across the two interfaces, leading to different parts of the passage being visible for the same item.) In the Auto condition, only a portion of the last paragraph was visible on screen when this item was selected (which contained underlined portions for the last two questions of the passage). Item 29 also required the examinee to choose the correct tense for an underlined word, and the correct response was "No Change." So it is possible that Auto computer examinees were helped on Item 30 by the passage layout, if they responded correctly to Item 29.

The findings for these items suggest that there are many different content and interface factors that could influence how an examinee responds to an item. These factors could include interface design features, features of specific questions, where the underlined portion occurs relative to response options, where page breaks occur on the paper version, features in adjacent items, and examinee tendencies themselves. Some of these factors might be controllable through the interface design. Others may depend upon a combination of individual circumstances, which might not always be controllable. If similar evaluations were conducted for all items in the test, it would probably give more insight into understanding what factors can affect examinee performance, and whether any additional modifications to the interface could help control those factors.

Reading

Plots of individual item p-value differences with error bands (computer p-value – paper p-value ± 2 standard errors) across paper and computer conditions are given in Figures 4-5. Each passage is separated by a vertical line in the plots. Figure 4 shows the computer – paper p-value differences for the Page and Scroll computer conditions from Comparability 2. Figure 5 shows the computer – paper p-value differences for Comparability 1 and for a baseline comparison based on two mutually exclusive random samples of examinees who took the items used in the comparability studies on paper as part of an equating study. (The form used for Reading in the Comparability 1 and Comparability 2 was previously administered in an equating study.) The computer and paper samples for the baseline comparison were created in the same manner as described for English. For all plots, a positive difference indicates that the item was easier on computer than paper. Table 5 gives the number (and percent) of Reading items for which the error bands do not surround zero in the Comparability 1 and Comparability 2 studies.

Results for Comparability 1 show zero items favoring computer examinees and 12 items (30%) favoring paper examinees. Results for the Page condition of Comparability 2 show four items (10%) favoring computer examinees and nine items (23%) favoring paper examinees. Results for the Scroll condition of Comparability 2 show three items (8%) favoring computer examinees and eight items (20%) favoring paper examinees. For the Reading baseline comparison, zero items showed a significant performance difference across groups.

Figure 4 shows a trend favoring paper examinees in the beginning of the test, for both the Page and Scroll conditions in Comparability 2. For about the last third of the test, there appeared

to be a trend for the test to favor computer examinees, and more so for the Page condition than the Scroll condition. Of the first 16 items, nine favored paper examinees in the Page condition and zero favored computer examinees. In the Scroll condition, eight of the first 16 items favored paper examinees and zero favored computer examinees. In Comparability 1, five of the first 16 items favored paper examinees and zero favored computer examinees. Of the last 13 items, four significantly favored computer examinees in the Page condition and zero favored paper examinees. In the Scroll condition, three of the last 13 items significantly favored computer examinees and zero favored paper examinees. In Comparability 1, zero of the last 13 items significantly favored computer examinees and five significantly favored paper examinees.

With respect to Interface 2 relative to Interface 1, the results seem to indicate fewer items favoring paper examinees overall, more items favoring paper examinees at the beginning of the exam, and more items favoring computer examinees at the end of the test. The trend toward favoring computer examinees toward the end of the exam might be attributable to greater ease of responding when speeded response behavior occurred at the end of the test. Results for Comparability 2 suggest there was some speeded response behavior. On average, Page examinees spent 51.3 seconds per item on the first 15 items, and 31.6 seconds per average on the last 15 items (51.6 and 31.6 seconds for Scroll examinees, respectively). As posed for English, the "no-bubble effect" might explain the trend favoring computer examinees toward the end of the test. Once examinees start to rush to complete the exam, it might be advantageous to take the test on computer rather than on paper, because of the ease in responding and moving quickly through items, and a greater ability to focus on the question at hand without being distracted by extraneous information. The favoring of paper examinees at the end in Comparability 1 could have occurred because in Interface 1 it was less easy for computer examinees to respond quickly once speeded response behavior began.

At face value, it is unclear why there would be more items favoring paper examinees at the beginning of the test under the new interface than under the old. The changes made to the interface were designed to improve the speed with which examinees could navigate throughout the passage (in the Scroll condition), to facilitate the occurrence of positional memory for specific content in the passage (in the Page condition), and to improve pre-test training on how to navigate. Figure 6 shows p-value difference plots for Reading items 4, 6, 9, 24, 25, 29, and 30. Each item plot contains the computer-paper p-value difference ± 2 standard errors for

Comparability 1, for the Scroll and Page condition from Comparability 2, and for the baseline comparison. Hypotheses explaining performance differences observed in Comparability 1 will be discussed here for a handful of items, and new findings will be discussed relative to the original hypotheses. Again, reasons why results occurred are speculated, but actual causes of the performance differences cannot be determined with any certainty.

Item 4. This item referred to a specific part of the passage, but no line reference was given, so there was undirected scrolling or paging to find the information. Results for Comparability 1 significantly favored paper examinees. It was hypothesized that the item required a lot of scrolling to find the appropriate reference in the text and that as such, the information was difficult to find in the passage. In Comparability 2, scrolling speed was increased, pre-test training was better on scrolling for the Scroll condition, and a Paging condition was added. Including the Page condition might have increased the likelihood that an examinee using that interface might remember where in the passage the information occurred. However, the results for both the Scroll and Page condition significantly favored paper examinees in Comparability 2. The percentages responding to each option show that computer examinees in both Comparability 1 and Comparability 2 were distracted by another response option that contained information that was given in the passage, but that was not the correct response to the question.

Because of the navigation necessary to locate the right answer in the passage, it is possible that computer examinees were more likely to stop reviewing the passage after finding a correct-looking option than were paper examinees, if they located it in the passage before finding the correct answer. Having found a correct-looking response, they might have selected it without looking further to see whether any of the other response options were contained in the passage. The extra navigation required to check out all response options may have prohibited them from continued checking after selecting an initial response.

Item 6. This item assumed a global understanding of the passage. The answer was not stated directly in the passage, but rather, the reader had to infer from the passage the correct response. Results for Comparability 1 did not favor either paper or computer examinees, so this item was expected to be neutral in Comparability 2 also. However, results for both the Scroll and Page condition significantly favored paper examinees. On average, the computer examinees did not spend any more time on this item than other items, so it does not appear that they were

looking for an answer that they could not find. They did, however, choose an incorrect answer more frequently than paper examinees did. Again, the computer examinees in Comparability appeared distracted by a reasonable looking option that could be inferred as correct by someone who did not read the passage carefully. It is not clear why Comparability 2 computer examinees were negatively affected while Comparability 1 examinees were not, unless the improved navigation and increased navigation speed caused examinees to read the passage less carefully under the revised interface.

Item 9. This item was difficult. The answer was not explicitly stated in the passage and required undirected scrolling to find relevant information in the passage. Results for Comparability 1 significantly favored paper examinees. It was expected that improved navigation in Interface 2 would decrease the mode effect in Comparability 2, however, the item still significantly favored paper examinees, for both the Scroll and Page conditions. In both studies, the computer examinees were distracted by one reasonable looking response option. Again, when the answer is not explicitly stated in the passage, or even when it is, computer examinees might be more inclined to select an incorrect option because it corresponds to information that they find first in the passage, without looking further to find information confirming the correct response option.

Item 24. This item referred examinees to a specific line in the passage and asked the meaning of the term "blue" in the referenced line. Results for Comparability 1 significantly favored paper examinees. It was hypothesized that the different content of the referenced line (caused by different line breaks across paper and computer modes) caused this performance difference. The line in the computer presentation contained both the word "blue" and "blues" which could have been confusing to computer examinees, whereas the line in the booklet presentation contained only the word "blue." In Interface 2, the line breaks were made the same across computer and paper modes, so the content of referenced lines was identical. We expected that there would be no performance difference for this item in Comparability 2, and there was not for either the Page or Scroll version.

Item 25. This item referred examinees to a specific line in the passage. Item 24 also referred examinees to a specific line later in the passage, so that all examinees had to move from the one location to the other to find the referenced line. Results for Comparability 1 significantly favored paper examinees. It was hypothesized that navigation difficulties and slow scrolling

speed interfered with computer examinees' performance on this item. In Comparability 2, there was no significant difference in performance across paper and computer modes, for both the Scroll and Page condition, although the Scroll condition showed an advantage for paper examinees.

Item 29. This item was a very difficult item. The answer for this item was contained in the second to last paragraph of the passage. Results for Comparability 1 significantly favored paper examinees. It was hypothesized that this item required a lot of navigation through the passage to find the answer in the passage, and that computer examinees had difficulty navigating through the passage. In Comparability 2, navigation speed was increased (through an increased scrolling speed in the Scroll condition, the use of paging in the Page condition, and better training on how to navigate through the passage). Results for both the Scroll and Page conditions showed no significant difference, although the Scroll condition showed an advantage for paper examinees.

Item 30. This item was also a very difficult item that required a global understanding of the passage. No explicit answer was stated in the passage. Results for Comparability 1 favored computer examinees, although not significantly. If this item were consistent with other items with similar characteristics, we would expect that it would have favored paper examinees because the answer was not explicitly stated and the examinee had to navigate to find the relevant information to answer the question. However, it was hypothesized that computer examinees were advantaged by the passage layout and their responses to the previous question. The correct response for this item referred to the "blues" and was the only response option to contain the word blues. The paragraph that contained the answer to Item 29 also referred to the blues. Computer examinees that had that paragraph visible on screen from answering Item 29 might have been more inclined to select the response option that also referred to the blues, because the word blues was already visible in the passage window. Results for Comparability 2 showed the same trend as Comparability 1, for both the Scroll and Page condition. Computer examinees were favored in both conditions, and the difference was significant for the Page condition.

As these examples indicate, the item-level results for Reading are still somewhat inconclusive as to why items early in the test favored paper examinees so heavily. Multiple factors may have contributed to the mode differences. The results do suggest, however, that

Comparability 2 computer examinees behaved differently from paper examinees, particularly early in the test. There might have been some learning effects early in the test, where examinees were still learning how to navigate and take the test on computer, so that their test-taking strategies differed from paper examinees. It does appear that computer examinees were more distracted by feasible looking response options, particularly under Interface 2. They may have been more likely to stop reviewing the passage once they found an acceptable looking response, without checking the other response options. Improving navigation speed might possibly have facilitated this behavior under Interface 2, if examinees read the passage less carefully. Further evaluation of the item and interface design features for other items may raise some new explanations for observed performance differences across modes and interface variations.

Whereas it appears that increasing navigation speed and allowing for the occurrence of positional memory (in the paging condition), along with the no-bubble effect might have helped improved overall computer completion rates across the two interfaces, those changes might not accommodate for the fact that the entire passage is not visible at once. The density of the text may make it more difficult for the computer examinees relative to the paper examinees, particularly early in the test. It is possible that adding more white space around the text would help computer examinees find information more easily, but that would require longer passages, which could offset the gain in adding white space, particularly for the scrolling variation.

Science Reasoning

Plots of individual item p-value differences with error bands (computer p-value – paper p-value ± 2 standard errors) across paper and computer conditions are given in Figures 7-8. Each passage is separated by a vertical line in the plots. Figure 7 shows the computer – paper p-value differences for the Page and Scroll computer conditions from Comparability 2. Figure 8 shows the computer – paper p-value differences for Comparability 1 and for a baseline comparison based on two mutually exclusive random samples of examinees who took the items used in the comparability studies on paper as part of an equating study. (The form used for Science Reasoning in Comparability 1 and Comparability 2 was previously administered in an equating study.) The computer and paper samples for the baseline comparison were created in the same manner as discussed for English. In all plots, a positive difference indicates that the item was easier on computer than paper. Table 5 gives the number (and percent) of Science Reasoning items for which the error bands do not surround zero in Comparability 1 and Comparability 2.

Results for Comparability 1 show six items (15%) significantly favoring computer examinees and seven items (18%) significantly favoring paper examinees. Results for the Page condition of Comparability 2 show 13 items (33%) significantly favoring computer examinees and five items (13%) significantly favoring paper examinees. Results for the Scroll condition of Comparability 2 show 10 items (25%) significantly favoring computer examinees and five items (13%) significantly favoring paper examinees. For the Science Reasoning baseline comparison, three items (8%) showed a significant performance difference across groups. These are likely due to chance.

Figure 8 shows that in Comparability 1, some trends in performance differences occurred within certain passages of the test. All items in the last passage significantly favored paper examinees, whereas Passage 4 strongly favored computer examinees (4 of the 6 items showed significant differences). In Passage 1, Item 2 and Item 3 significantly favored computer examinees, while Item 6 favored paper examinees. Within the remaining three passages, there was no apparent trend favoring either computer or paper examinees. The effect for the last passage might be attributable to speeded response behavior and non-completion factors, which could have disadvantaged computer examinees. Results for Comparability 2 suggest there was still some speeded response behavior. On average, Page examinees spent 54.8 seconds on the first 15 items, and 30.3 seconds on the last 15 items (56.3 and 30.9 seconds for Scroll examinees, respectively). Taking into account the different level of ability across the two studies, there likely was a greater rate of non-completion for Comparability 1 than Comparability 2.

For Comparability 2, findings were a bit different within passages. The last passage was neutral for both the Page and Scroll conditions (with the exception of Item 37 significantly favoring paper examinees in the Page condition). This suggests that the interface changes might have removed some of the factors that caused computer examinees to be disadvantaged in Comparability 1 once speeded response behavior began. Within Passages 3-6, there was a trend for items to favor computer examinees. For the Page condition, 11 of the 22 items significantly favored computer examinees. For the Scroll condition, seven of the 22 items significantly favored computer examinees. For Comparability 1, four of the 22 items favored computer examinees (all in Passage 4). For these passages, much like observed in the English test, this trend favoring computer examinees might be attributable to the "focus effect." Particularly for Science Reasoning, it might be beneficial to computer examinees to be able to view only the

relevant graphic on screen, and remove extraneous information contained in the rest of the test from the screen. It is unclear, however, why the last passage did not also favor computer examinees.

Across the first two passages, there appeared more of a trend for items to favor paper examinees in Comparability 2. In both the Page and Scroll conditions, four of the 12 items significantly favored paper examinees (versus one significantly favoring paper in Comparability 1). Item 2 and Item 3 still significantly favored computer examinees (as observed in Comparability 1 also), along with Item 12 in the Scroll condition. What is happening with Item 2 and Item 3 is unclear. There is possibly some content factor that causes them to strongly favor computer examinees. For the rest of the items, there might have been learning effects early in the test, where examinees were still learning how to navigate and take the test on computer. Science Reasoning would benefit from a thorough evaluation of item and interface design features, such as performed for the English and Reading items discussed earlier. Individual items are not discussed for Science Reasoning, because no hypotheses were developed for Comparability 1 performance differences. It is possible that some of the performance differences can be accounted for through such an evaluation.

Item Parameter Differences

Item responses for the Comparability 2 paper and computer samples were calibrated using Bilog. The calibrations were conducted under the same conditions across the paper and computer samples. Not reached and omitted responses were treated as incorrect because that is how they are scored operationally for paper examinees. Correlations between paper and computer item parameters for Comparability 2 are given in Table 6. The b parameters were very highly correlated. The a parameters were less highly correlated than the b parameters, but were still fairly highly correlated. The c parameters were moderately correlated across modes.

Correlations between the computer parameters and re-estimated computer parameters for Comparability 2 are given in Table 7, as a baseline comparison. For the re-estimated computer parameters, the computer parameters were used to generate item responses in a normally distributed sample of examinees of the same size as the original calibration sample. The simulated 0,1 responses were then calibrated under the same conditions as the original computer parameters. A comparison of the original parameters (treated as the "true" parameters) with the re-estimated parameters indicates how much difference we would expect between paper and

computer parameters simply due to estimation error in the calibration process. The correlations show there was some estimation error associated with the calibration process, mainly in the a and c parameters. The original computer and re-estimated computer parameter correlations were higher than the computer and paper correlations, suggesting that there were some mode effects contributing to the paper and computer parameter differences observed in Comparability 2.

Table 6. Correlation Between Computer and Paper Item Parameters for Comparability 2

		Correlation		
Test	Computer Condition	a	b	c
English	Auto	.81	.93	.70
	Semi	.85	.95	.66
Reading	Page	.82	.96	.65
	Scroll	.77	.93	.59
Science Reasoning	Page	.68	.93	.45
	Scroll	.80	.95	.78

Table 7. Correlation Between Original Computer and Re-estimated Computer Item Parameters for Comparability 2

		Correlation		
Test	Computer Condition	a	b	c
English	Auto	.87	.97	.80
	Semi	.90	.97	.79
Reading	Page	.92	.99	.76
	Scroll	.78	.99	.82
Science Reasoning	Page	.86	.99	.84
	Scroll	.90	.99	.82

Figures 9-14 plot the Comparability 2 computer and paper parameters against one another and the original computer and re-estimated computer parameters against one another, for Reading and Science Reasoning. The plots of the computer vs. paper parameters showed more spread than the plots of the original computer vs. re-estimated computer parameters, again suggesting that there were some mode effects contributing to the parameter differences observed in Comparability 2.

Are the Parameter Differences Important?

Results for all content areas suggest that there were some significant mode effects contributing to performance differences. These mode effects resulted in item parameters that differed across paper and computer calibration samples. As such, this raises the question of whether item parameters calibrated from a paper administration can be used for operational computer administrations. Item parameters may differ across calibration samples, but if they are not of a magnitude to adversely affect an examinee's score if he or she were to take the items in one mode versus another, then we may not need to be concerned about the differences.

A simulation was performed to examine the effect on test-retest reliability of using paper parameters to administer items and score responses for computer examinees. All analyses used Comparability 2 parameters that were cloned to create item pools eight times the size of the original form (i.e., the item parameters were repeated eight times). This simulation assumes that the item parameters from a form are representative of what the item parameters would be in an operational pool. Thetas for 10,000 examinees were generated from a N(0,1) distribution and each examinee was administered two adaptive tests using maximum information item selection and exposure control parameters computed using the Sympson-Hetter algorithm. The Pearson product-moment correlation was computed between the final ability estimates from the two tests. Simulations were conducted for Reading Page, Reading Scroll, Science Reasoning Page, and Science Reasoning Scroll, for test lengths between 10 and 40. Reading and Science Reasoning were used because the test form administered in the comparability studies was an intact form that had been administered operationally, and so met the specifications for paper forms.

For each content area, three different conditions were simulated, labeled "Comp(T)/Comp(T)," "Comp(T)/Comp(E)," and "Comp(T)/Paper(E)." The three conditions differed in terms of the item parameters used to generate response, select items, and score responses. The conditions are summarized in Table 8.

Under the Comp(T)/Comp(T) condition, the computer parameters (treated as "true") were used to generate item responses, select items, and score responses. Items were selected based on information tables and exposure control parameters computed from the computer parameters, responses were generated based on the computer parameters, and intermediate and final ability estimates were based on the computer parameters. This represents what would happen operationally if the true computer parameters were used in a computer administration.

Table 8. Parameters Used to Generate Responses, Select Items, and Score Responses for the Three Simulation Conditions.

Condition	Generate Responses	Select Items	Score Responses
Comp(T)/Comp(T)	Computer	Computer	Computer
Comp(T)/Comp(E)	Computer	Re-estimated Computer	Re-estimated Computer
Comp(T)/Paper(E)	Computer	Paper	Paper

Under the Comp(T)/Comp(E) condition, the computer (true) parameters were used to generate item responses, and the re-estimated computer parameters were used to select items and score responses. (The re-estimated computer parameters were those discussed earlier.) Items were selected based on information tables and exposure control parameters computed from the re-estimated computer parameters, responses were generated based on the original computer parameters, and intermediate and final ability estimates were based on the re-estimated computer parameters. This represents what would happen operationally if calibrated computer parameters were used in a computer administration. This would be the case operationally, as the true computer parameters would never be known. This condition provides a baseline comparison for how much the reliability results were affected simply by estimation error in the calibration process to obtain the computer parameters.

Under the Comp(T)/Paper(E) condition, the computer (true) parameters were used to generate item responses, but the paper parameters were used to select items and score responses. Items were selected based on information tables and exposure control parameters computed from the paper parameters, responses were generated based on the computer parameters, and intermediate and final ability estimates were based on the paper parameters. This represents what would happen operationally if the calibrated paper parameters were used in a computer administration.

A comparison of the results for the Comp(T)/Comp(E) condition and the Comp(T)/Paper(E) conditions relative to the Comp(T)/Comp(T) condition should indicate whether differences in reliability from the Comp(T)/Comp(T) condition are due simply to

estimation error in the calibration process, or are also due to the use of paper parameters in a computer administration.

The simulated test-retest reliabilities were modeled using a 3rd degree polynomial regression model. Figures 15-18 plot the predicted test-retest reliabilities by test length for the Comp(T)/Comp(T), Comp(T)/Comp(E), and Comp(T)/Paper(E) conditions for Reading Page, Reading Scroll, Science Page, and Science Scroll, respectively. The results for both Reading and Science Reasoning show some loss in score precision due to calibrating the item parameters used in the item selection and scoring. The results also suggest that we can expect some loss in precision above and beyond the loss due to calibrating the item parameters, if we were to use paper-calibrated parameters in a computer administration. The loss in precision due to use of paper parameters was greater for Science Reasoning than for Reading, which corresponds to the slightly greater spread observed for Science Reasoning than for Reading in the plots of the computer vs. paper parameters (Figures 9-14). The loss in reliability due to the use of paper parameters in a computer administration was greatest for test lengths of 10 for all conditions. As test length increased, the effect was diminished for all conditions but Science Page.

Operationally, if the computer test length is set to meet a target reliability, then a slightly longer test may be required if paper parameters are used in a computer administration. Target reliabilities were computed for each condition as the correlation between test-retest number right scores for 10,000 normally distributed examinees. Responses were generated to a fixed form using the original computer parameters from Comparability 2. This reliability represents the reliability of the test form in the calibration sample. Results from the polynomial model show that test lengths of 15, 15, 14, and 13 meet the target reliabilities of .85, .84, .84, and .85 for Read Page, Read Scroll, Science Page, and Science Scroll, respectively. The target lengths were determined by the Comp(T)/Comp(E) model, which represents use of calibrated computer parameters for computer administration.

In order to meet the target reliabilities for Read Page and Read Scroll, one additional item is needed if the paper parameters are used rather than the re-calibrated computer parameters in a computer administration. This corresponds to a test length that is approximately 7% longer. For Science Scroll, three additional items are needed if the paper parameters are used rather than the re-calibrated computer parameters. This corresponds to a test length that is approximately 23% longer. For Science Page, four additional items are needed if the paper parameters are used

rather than the re-calibrated computer parameters, if the paper parameters are used rather than the re-calibrated computer parameters. This corresponds to a test length that is approximately 29% longer.

These test lengths required to meet the target reliabilities may be shorter than a program would want to use operationally. With longer test lengths, fewer additional items might be needed to compensate for the loss in precision due to using paper parameters with a computer administration. This is the case for Read Page, Read Scroll, and Science Scroll. Results for Science Page suggest, however, that this may be dependent on the magnitude of the parameter differences across modes. Loss of precision due to parameter differences of a smaller magnitude might be controllable through longer tests, but once the parameter differences reach a certain magnitude, increasing test length might not have an effect.

Discussion

The findings from the two comparability studies, in conjunction with previous experienced garnered from reviews of test content, test booklets, computer interfaces, and interviews with examinees, suggest some answers to the three questions broached in this paper.

While some items showed no performance differences across administration modes, there were other items for which examinees did not respond in the same way across modes or interface variations. The analysis of the individual English and Reading items suggest that there are a variety of factors that could contribute to mode effects, and that each item presents a potentially unique set of circumstances that could cause different behaviors across modes. Intuition suggests that the more complex the test is, and the greater the differences in how passages and items are presented across modes, the greater the potential for performance differences across modes.

There may have been some sampling differences and some chance differences that affected the results of these studies, but in general, it appears that the changes made to the interface had some effect on computer examinees' performance on some items. Also, it is important to note that the effect was not always the intended effect. This suggests that examinees are sensitive and respond to how information is presented on computer, but not always in ways that are readily predictable. In some cases, the results appeared influenced by better pre-test training on how to use the functions necessary to take the test on computer, improved navigation speed and navigation capabilities, and making information about the test

session more readily available to examinees. While perhaps not technically part of the computer interface, all of these components contribute to the examinees' interaction with the interface, and should be considered in designing an interface and conducting computerized testing. Different results across interface variations also suggest that even within the same mode of administration, differences in how the test is presented could influence examinee behavior while testing. A seemingly subtle change such as aligning or not aligning the underlined portion in the English test with the top of the item can have a not-so-subtle effect on examinee behavior on some items. Thus, care also needs to be taken when implementing interface changes in an operational computerized testing program.

Although there were some significant p-value differences across modes, the magnitude of the p-value differences in general was not very large (i.e., $< \pm .10$ for almost all items). The results from the adaptive test simulations using the different item parameters showed a fairly small effect on the score precision when using paper parameters for item selection and scoring rather than computer parameters, particularly for longer test lengths. The simulation results could have been a little different had the not reached items been treated as not reached in the calibrations, rather than scored as incorrect. With not reached items scored as incorrect, items toward the end of the test could have appeared more difficult than they were in reality simply because fewer examinees completed them. If examinees did not get to an item because of slowness in responding due to the test interface rather than the content of the items, then the results might have appeared less reliable than they really were.

The results of the simulation suggest that for parameter differences of the magnitude observed in the Reading conditions and Science Scroll, item parameters calibrated from paper and pencil administrations could probably be used initially in a computer administration, with a longer test. If so, calibration sample sizes should be large to minimize the effect of estimation error. Then, when enough data from the computer administrations are available, the parameters could be re-calibrated and the operational test length shortened a little. The results for Science Page suggest, however, that the feasibility of doing so may depend on the particular test and the magnitude of the parameter differences across modes.

In all, the findings suggest that while performance differences do occur across modes, they may have a fairly small effect in practice. It is probably wise, however, to develop an understanding of the factors that can influence examinee behavior and design a computer

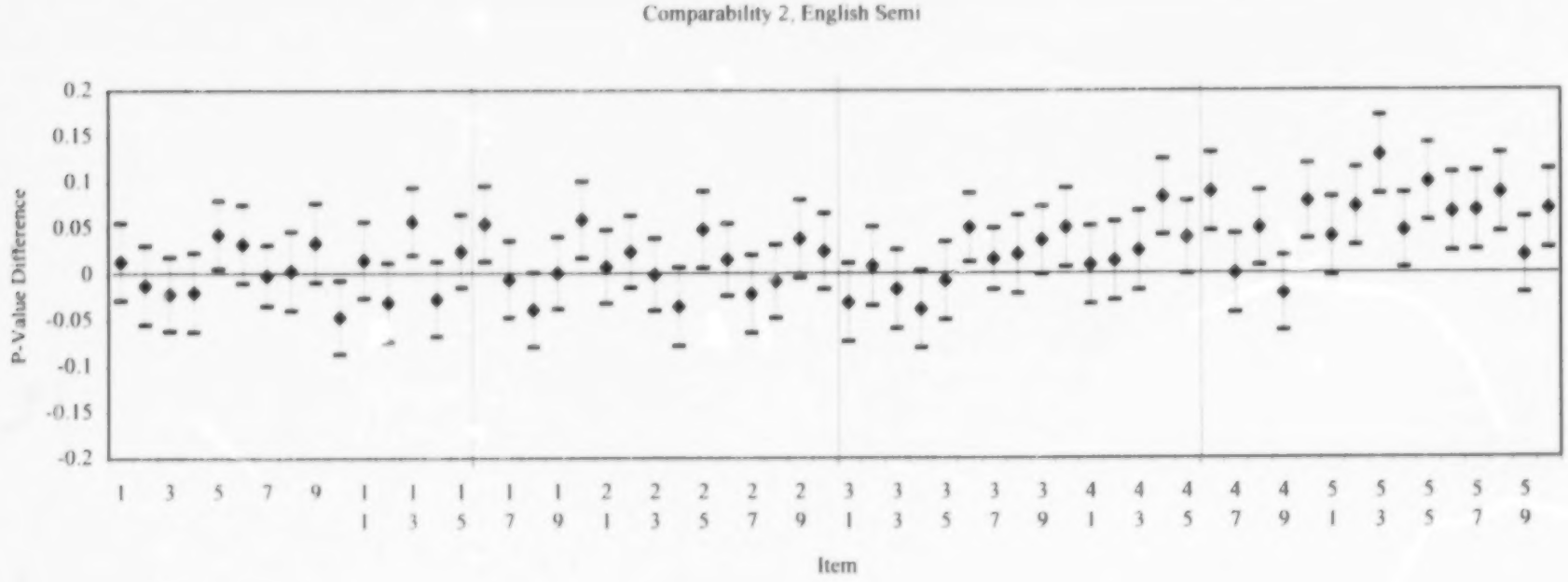
interface accordingly, to ensure that examinees are responding to test content rather than features inherent in presenting the test on computer. Information learned about how examinees interact with computer interface features through reviews of the type presented in this paper can help practitioners make decisions about how best to present complex tests via computer.

References

Parshall, C.G., Spray, J.A., Kalohn, J.C., & Davey, T. (2002). Practical considerations in computer-based testing. New York: Springer.

Pommerich, M. & Burden, T. (2000, April). From simulation to application: Examinees react to computerized testing. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Figure 1. English Computer - Paper P-Value Differences (+- 2 Standard Errors) for Comparability 2.



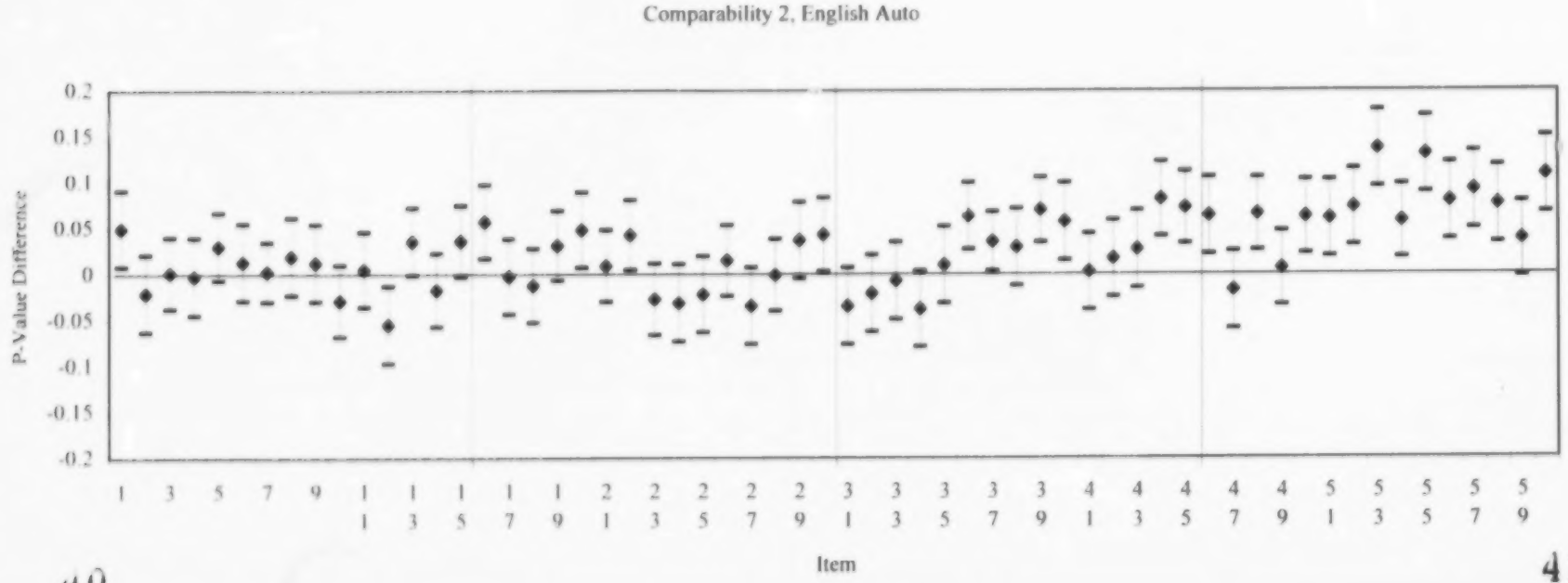
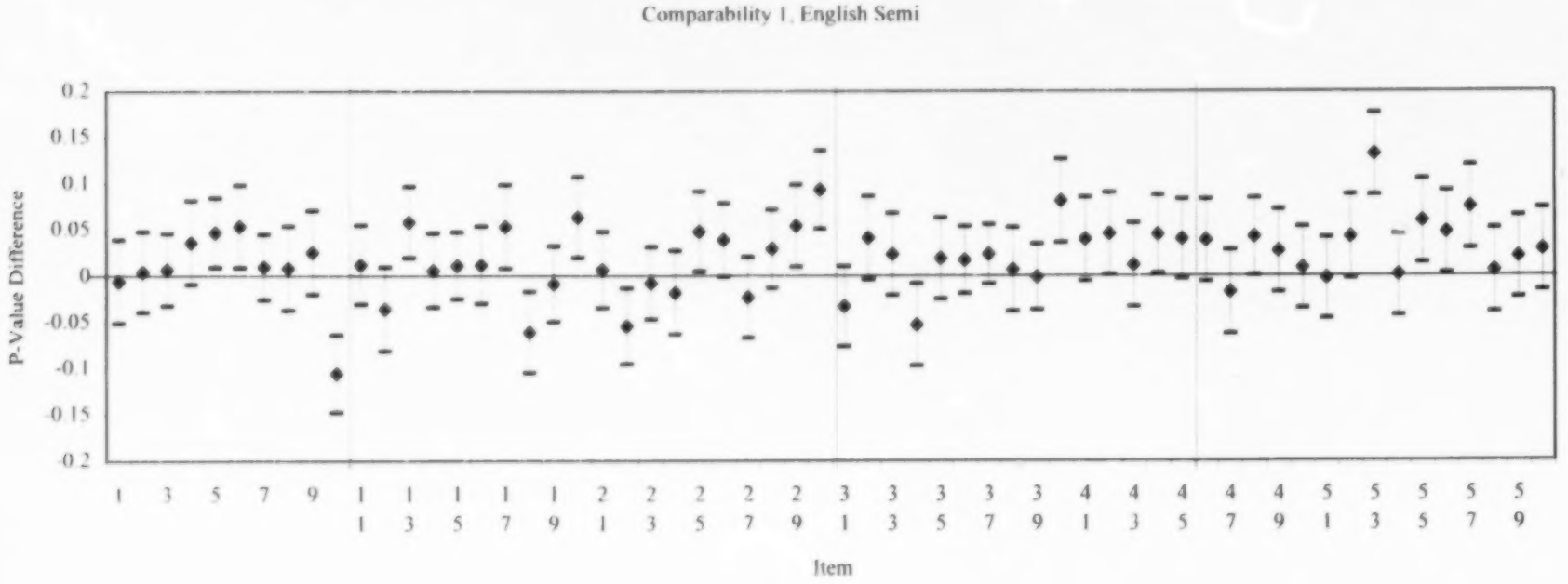


Figure 2. English Computer - Paper P-Value Differences (+- 2 Standard Errors) for Comparability 1 and a Baseline Comparision.



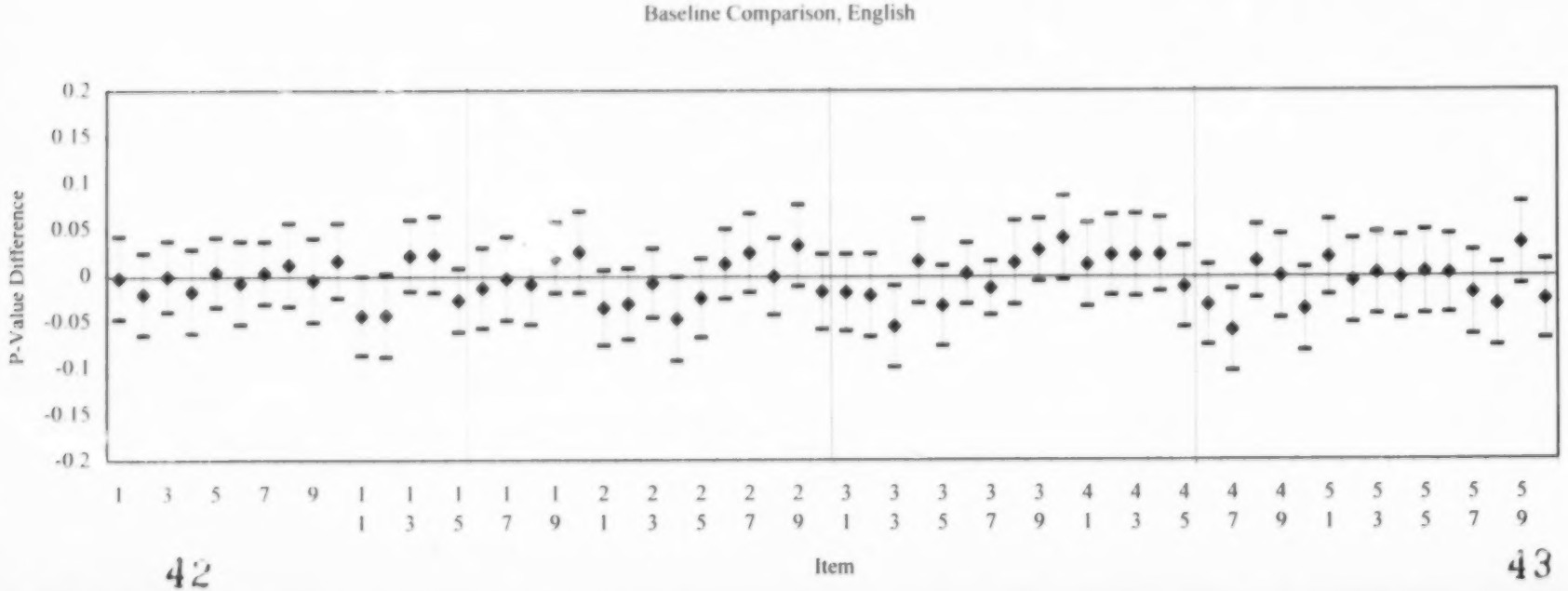
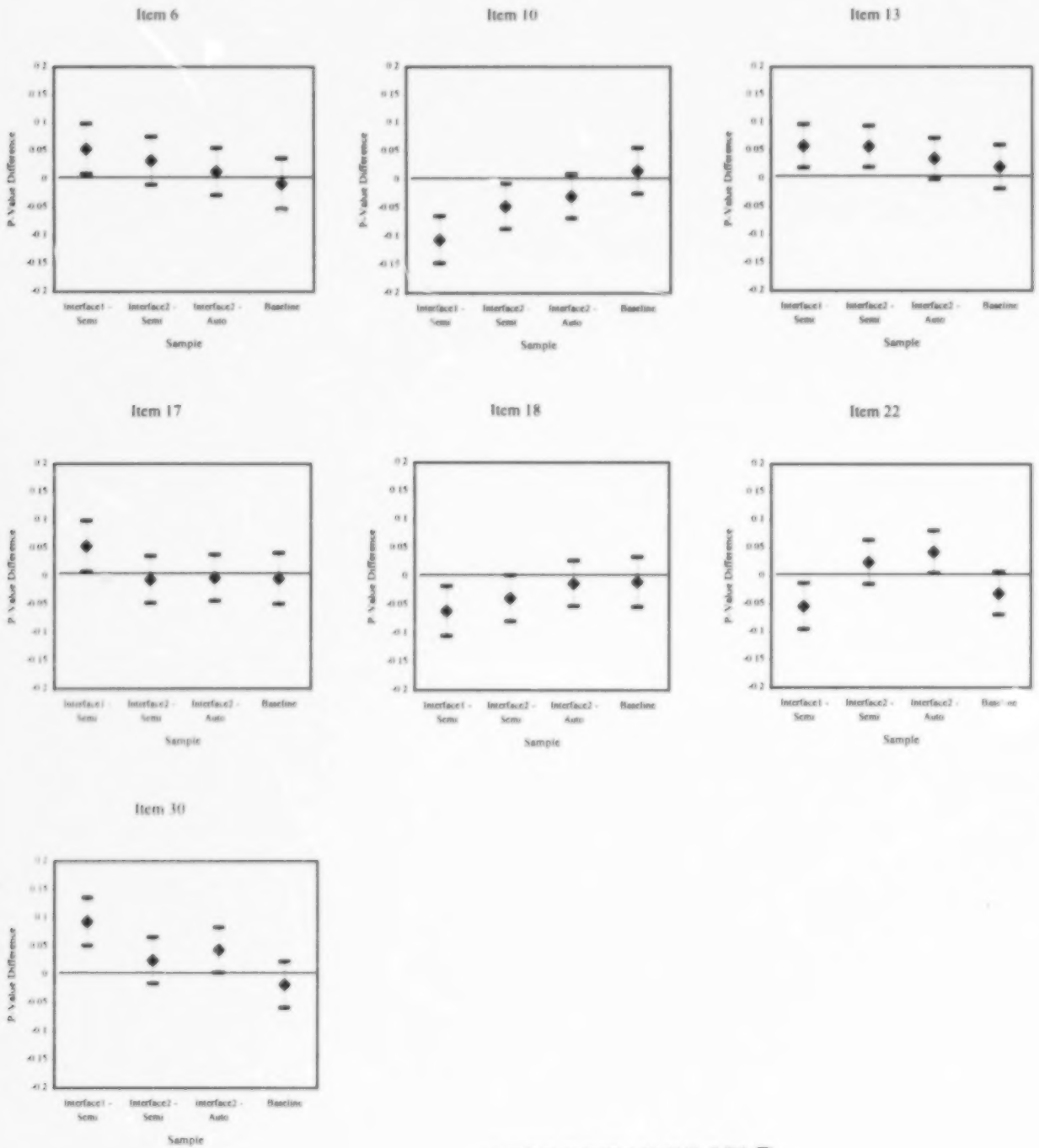
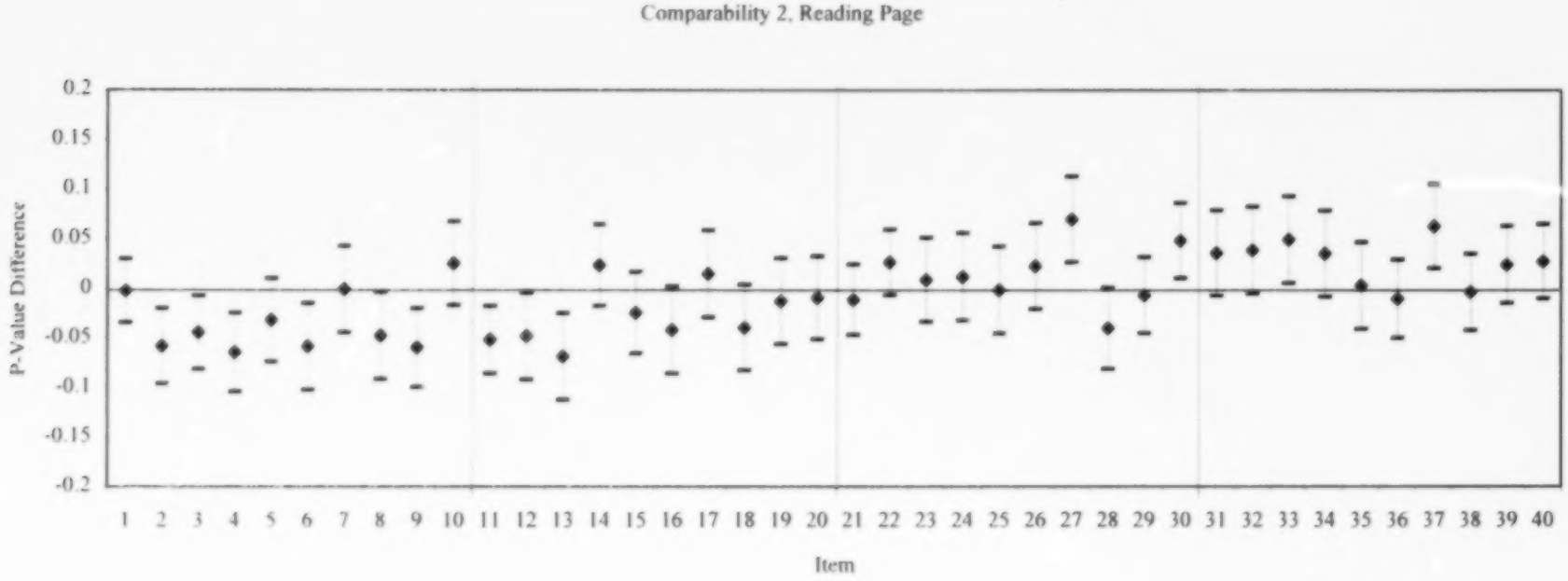


Figure 3. P-Value Plots for Individual English Items Summarized Over Comparability 1, Comparability 2, and a Baseline Comparison.



BEST COPY AVAILABLE

Figure 4. Reading Computer - Paper P-Value Differences (+- 2 Standard Errors) for Comparability 2.



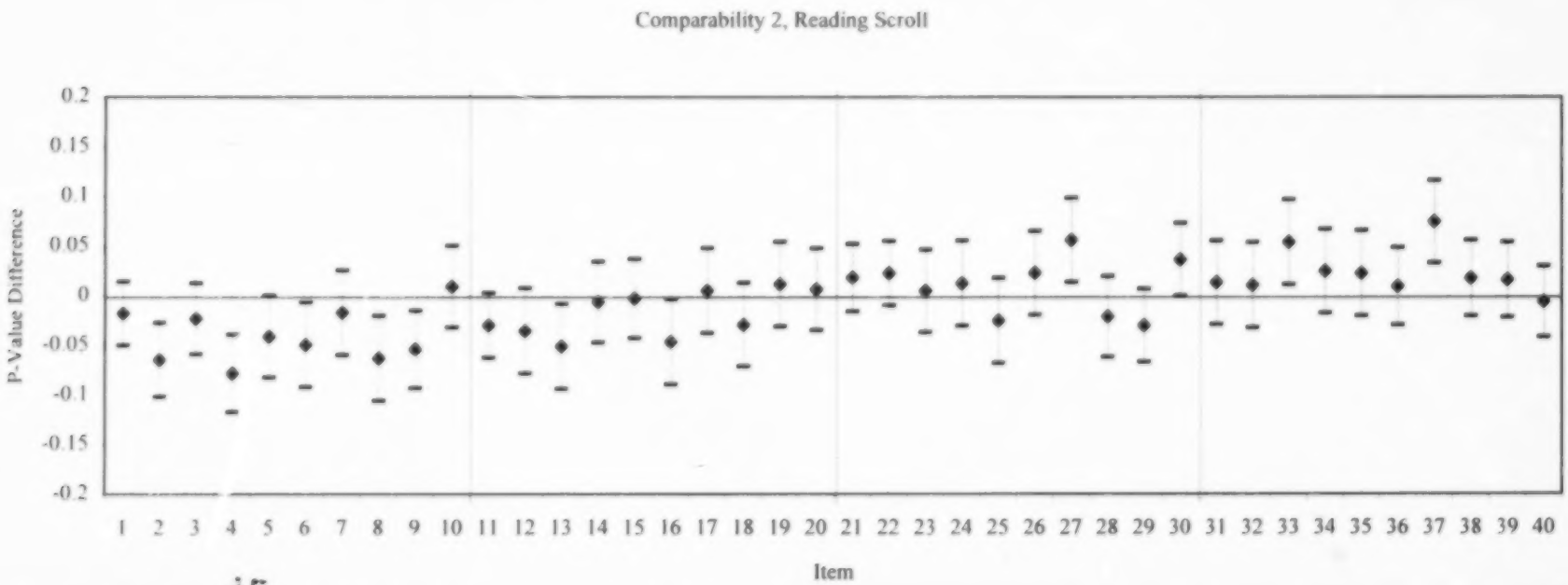
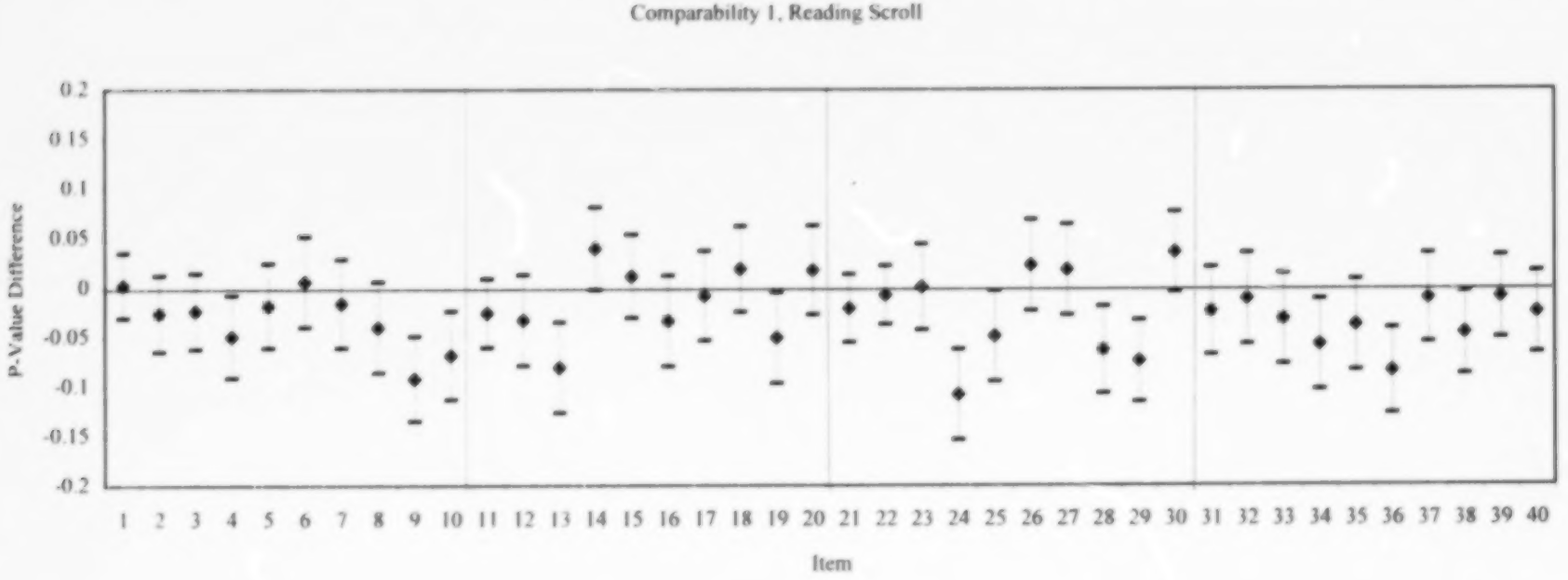


Figure 5. Reading Computer - Paper P-Value Differences (+- 2 Standard Errors) for Comparability 1 and a Baseline Comparison.



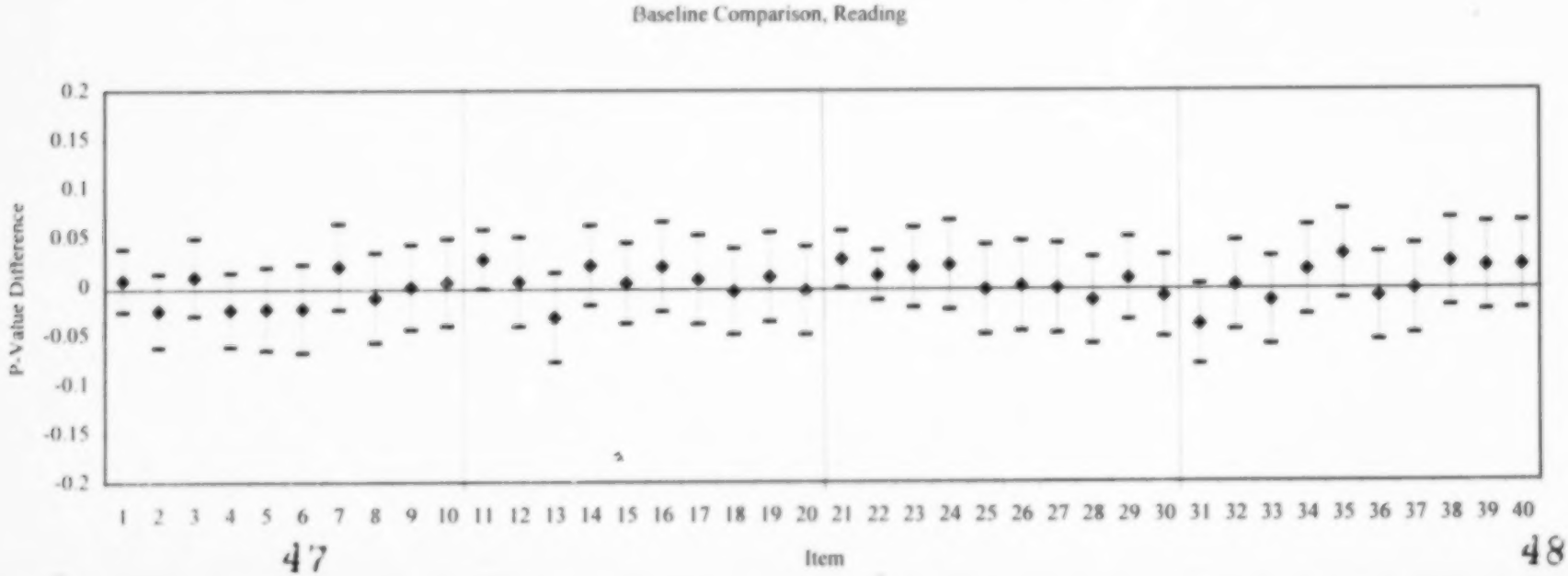
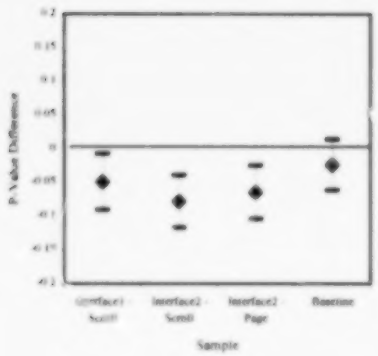
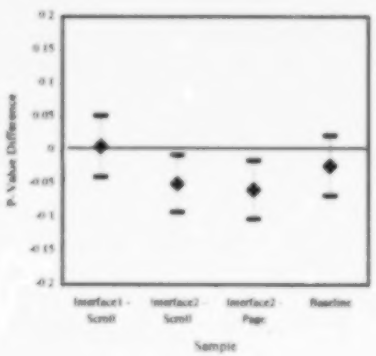


Figure 6. P-Value Plots for Individual Reading Items Summarized Over Comparability 1, Comparability 2, and a Baseline Comparison.

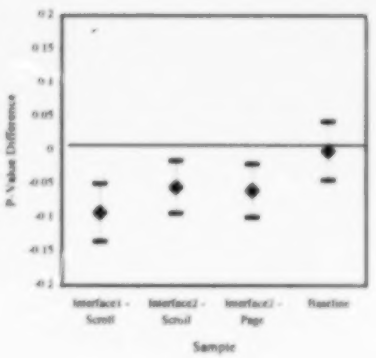
Item 4



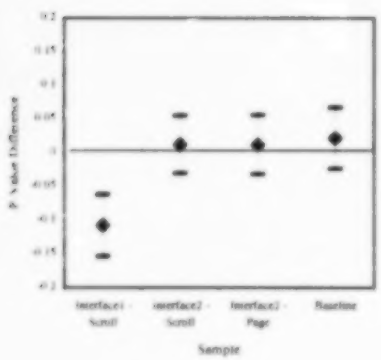
Item 6



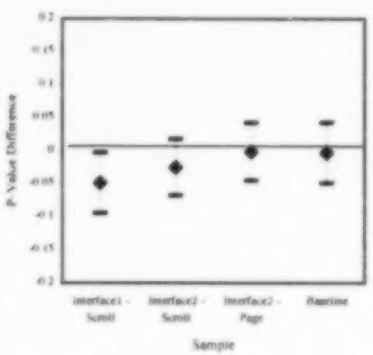
Item 9



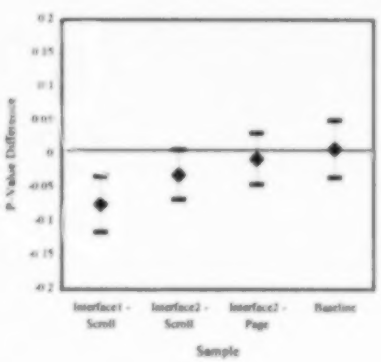
Item 24



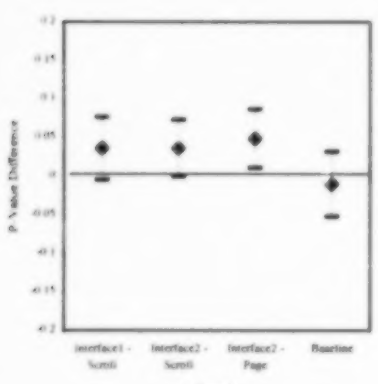
Item 25



Item 29

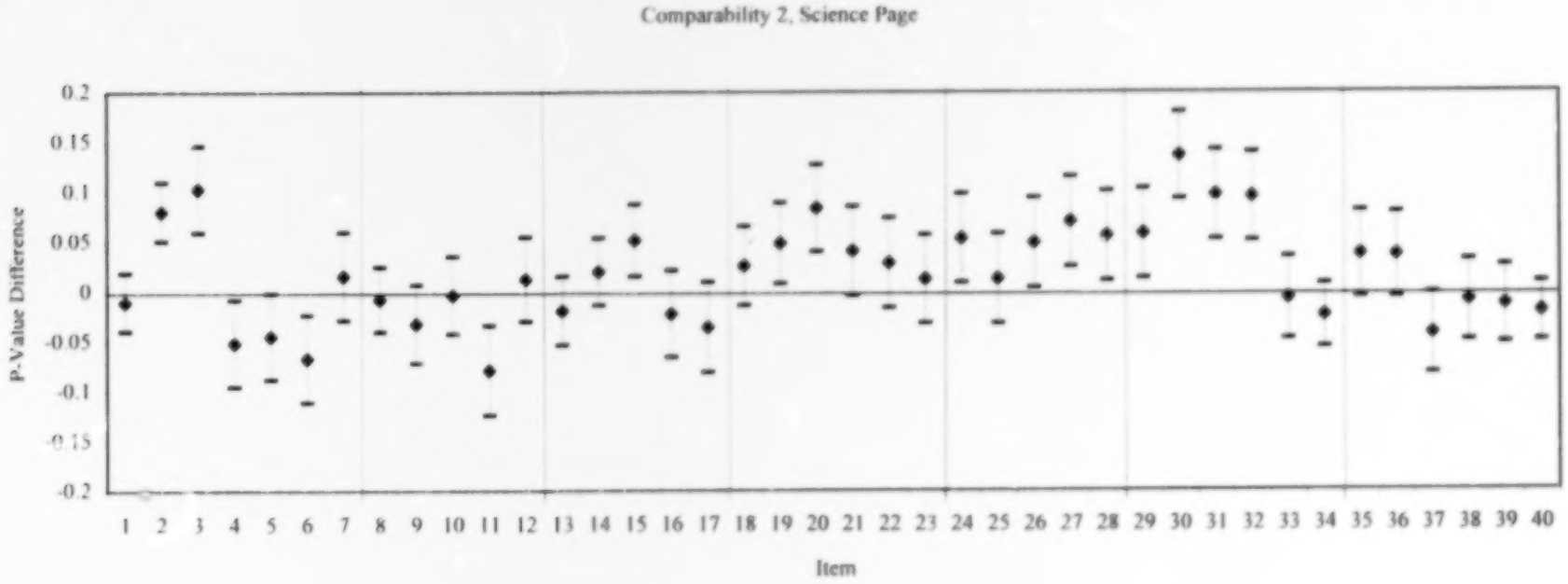


Item 30



BEST COPY AVAILABLE

Figure 7. Science Computer - Paper P-Value Differences (+- 2 Standard Errors) for Comparability 2



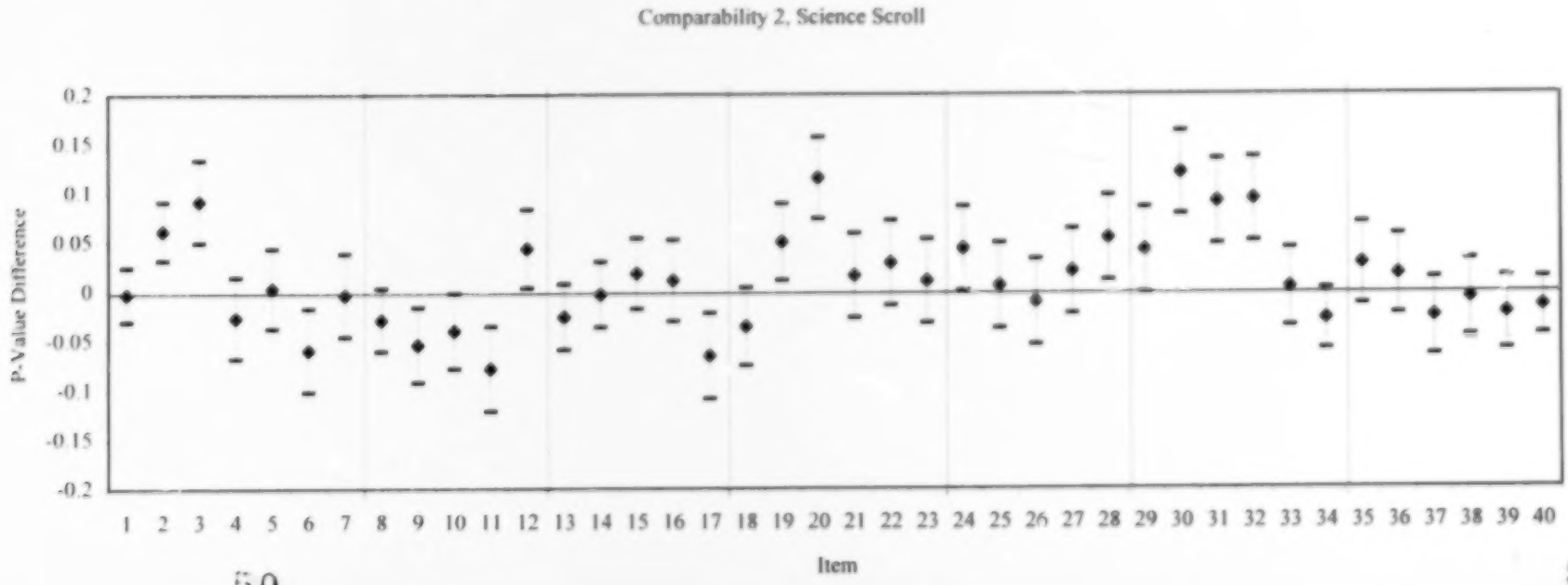
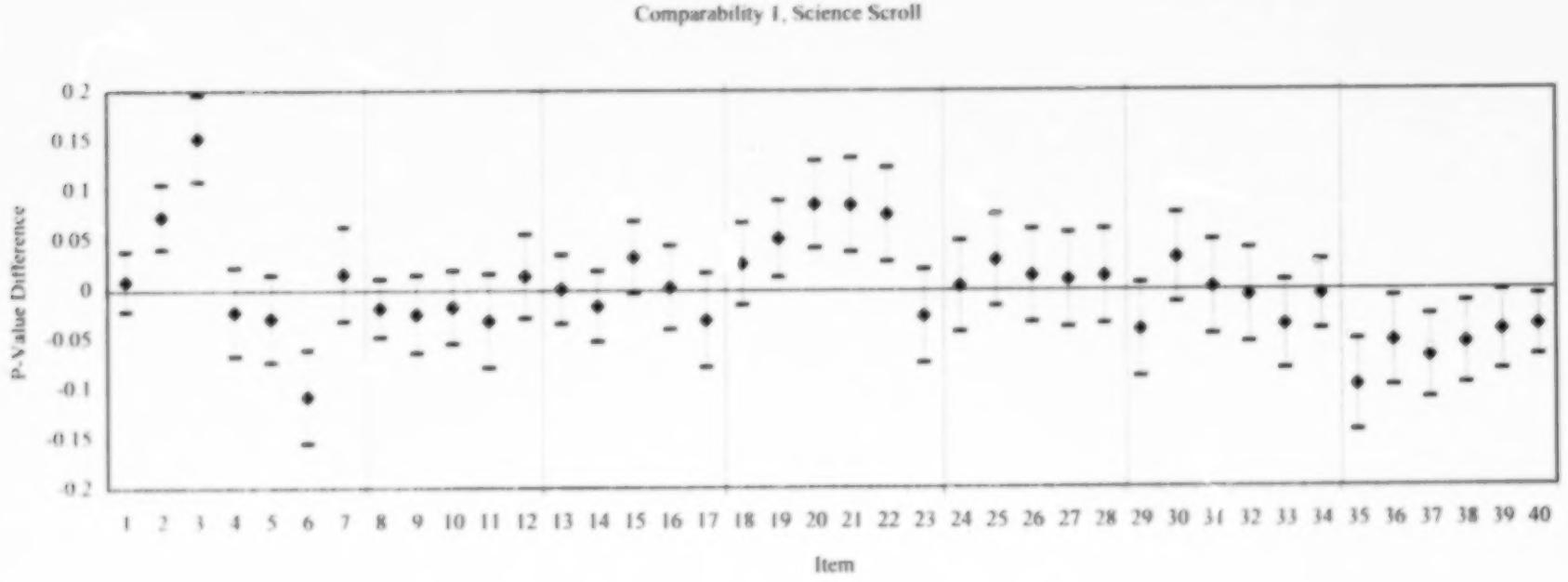


Figure 8. Science Computer - Paper P-Value Differences (+- 2 Standard Errors) for Comparability 1 a Baseline Comparison.



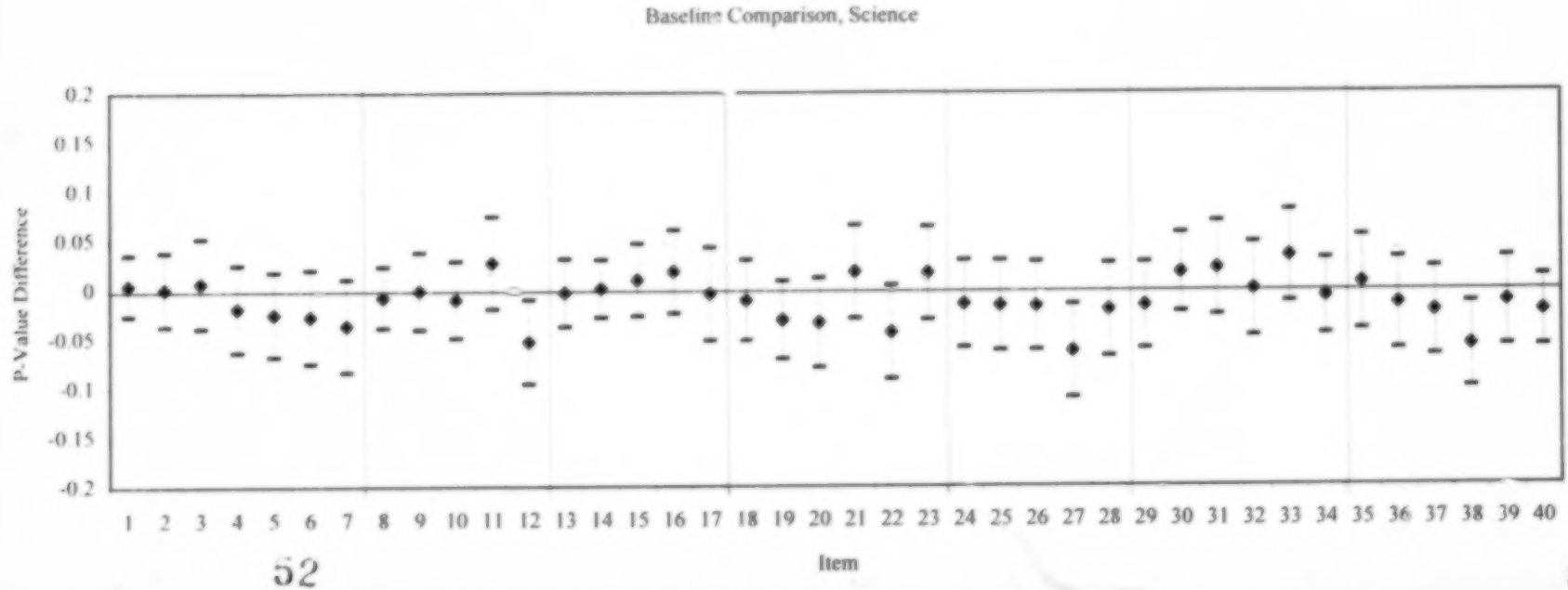
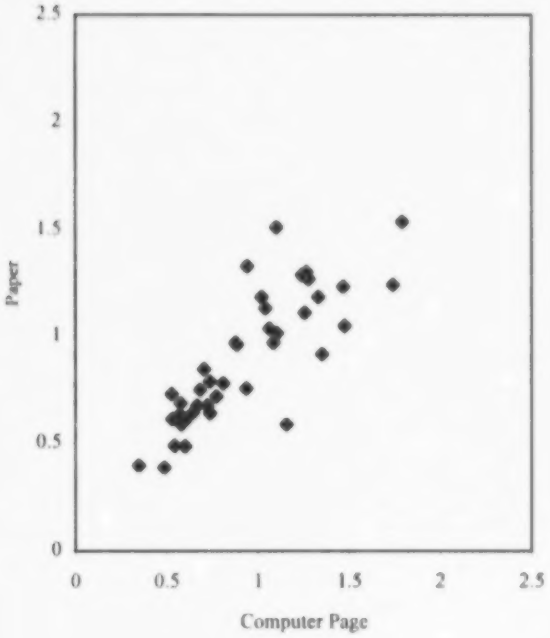
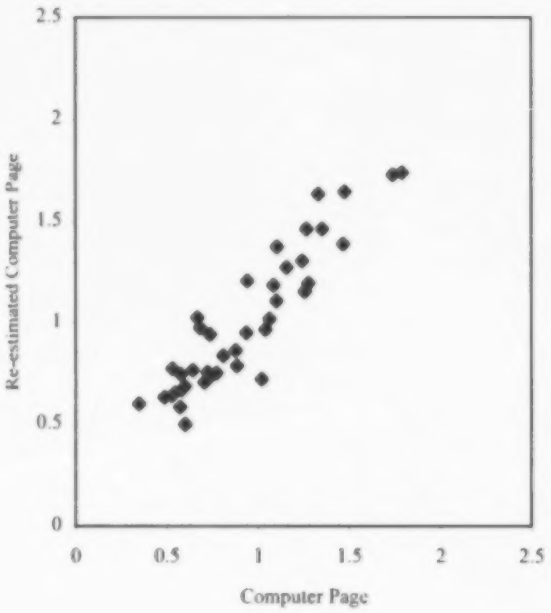


Figure 9. A Parameters for Comparability 2 Reading
(Computer Vs. Paper and Computer Vs. Re-estimated Computer).

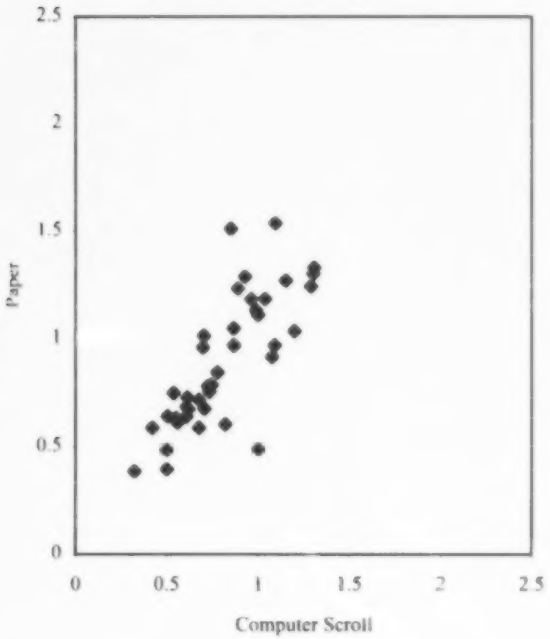
Computer Page Vs. Paper



Computer Page Vs. Re-estimated Computer Page



Computer Scroll Vs. Paper



Computer Scroll Vs. Re-estimated Computer Scroll

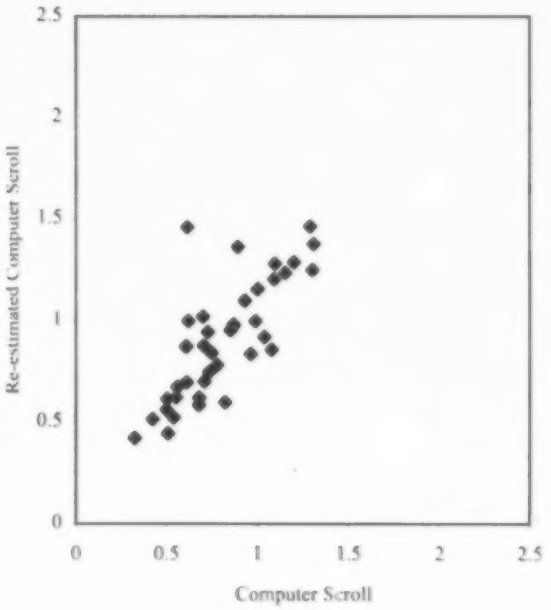
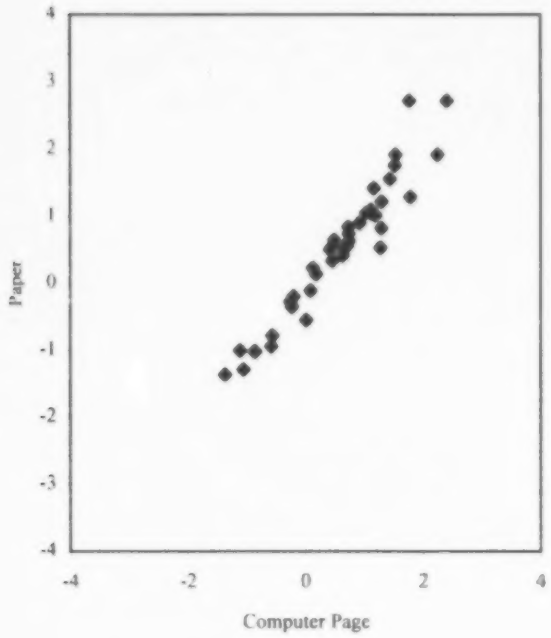
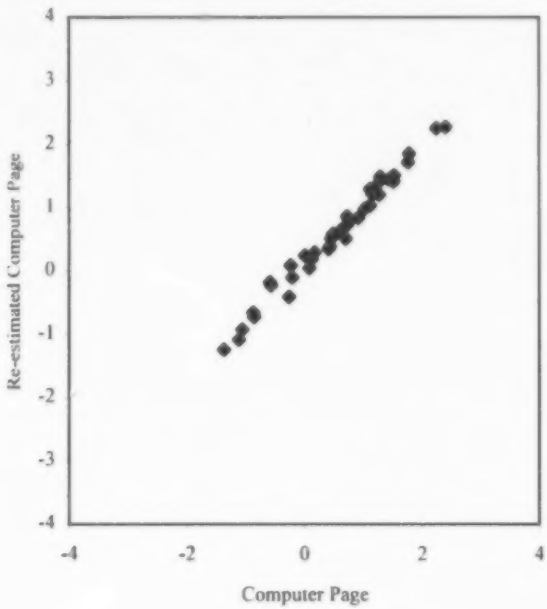


Figure 10. B Parameters for Comparability 2 Reading
(Computer Vs. Paper and Computer Vs. Re-estimated Computer).

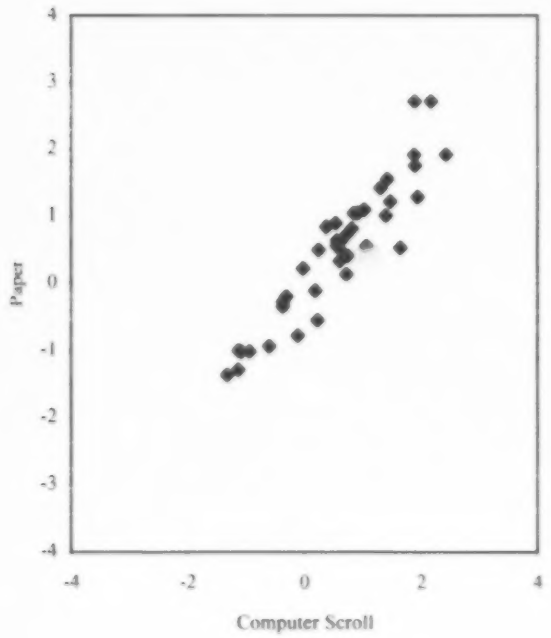
Computer Page Vs. Paper



Computer Page Vs. Re-estimated Computer Page



Computer Scroll Vs. Paper



Computer Scroll Vs.
Re-estimated Computer Scroll

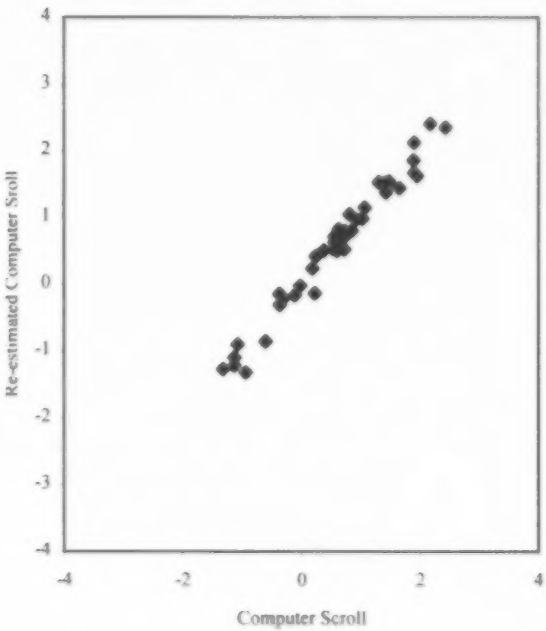
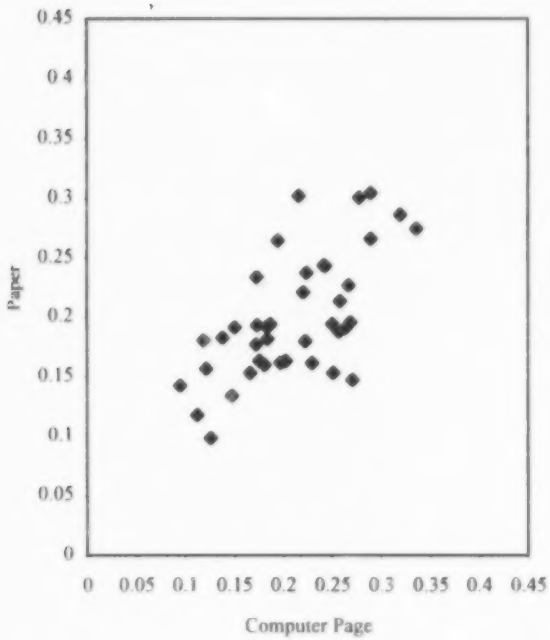
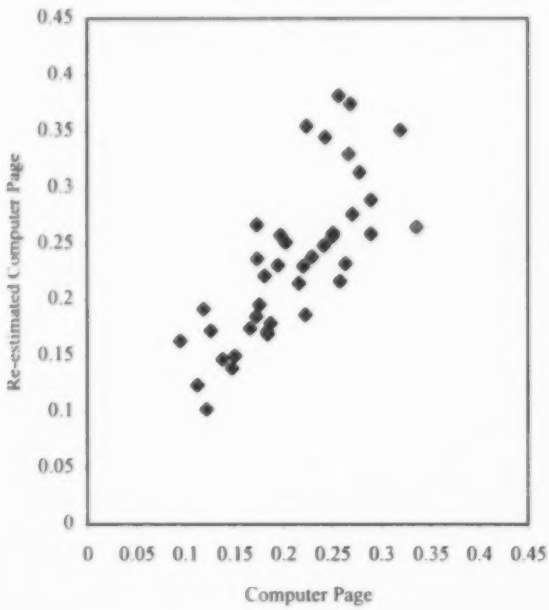


Figure 11. C Parameters for Comparability 2 Reading
(Computer Vs. Paper and Computer Vs. Re-estimated Computer).

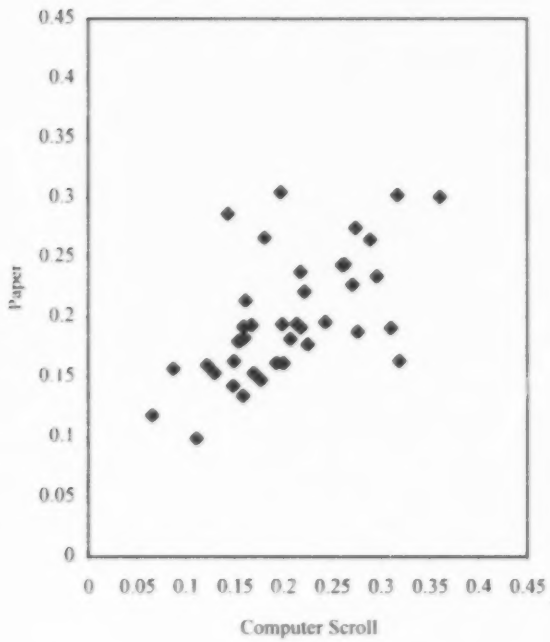
Computer Page Vs. Paper



Computer Page Vs. Re-estimated Computer Page



Computer Scroll Vs. Paper



Computer Scroll Vs. Re-estimated Computer Scroll

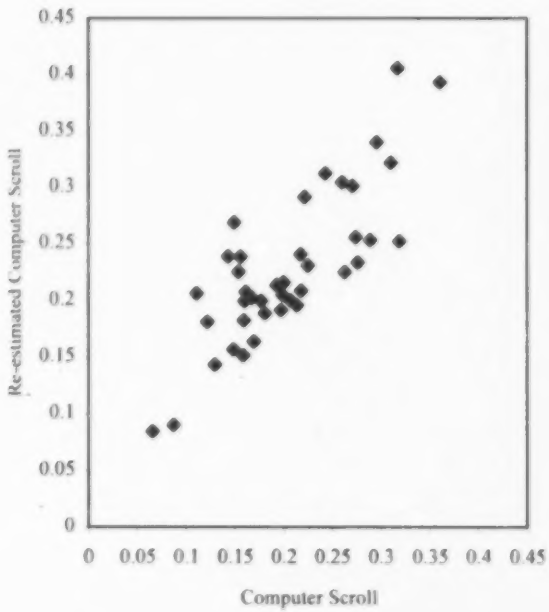
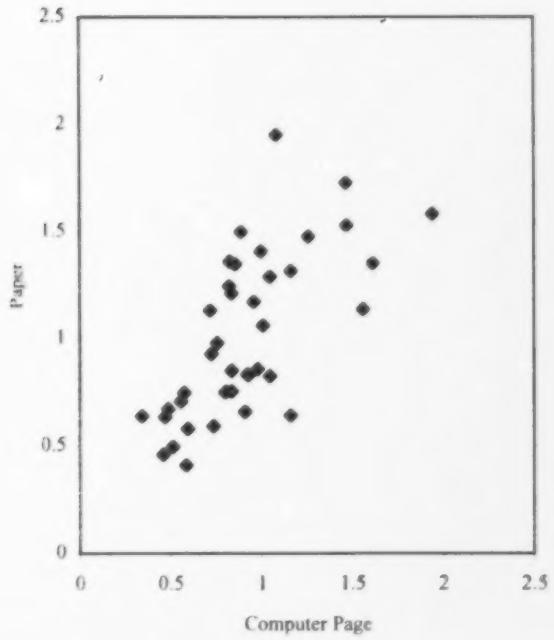
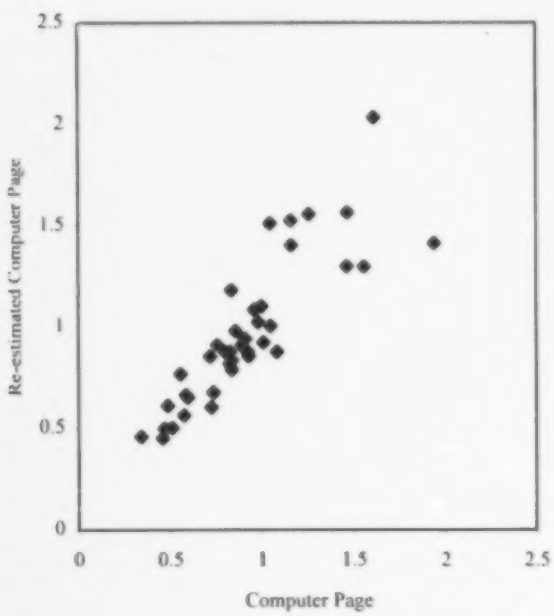


Figure 12. A Parameters for Comparability 2 Science Reasoning
(Computer Vs. Paper and Computer Vs. Re-estimated Computer).

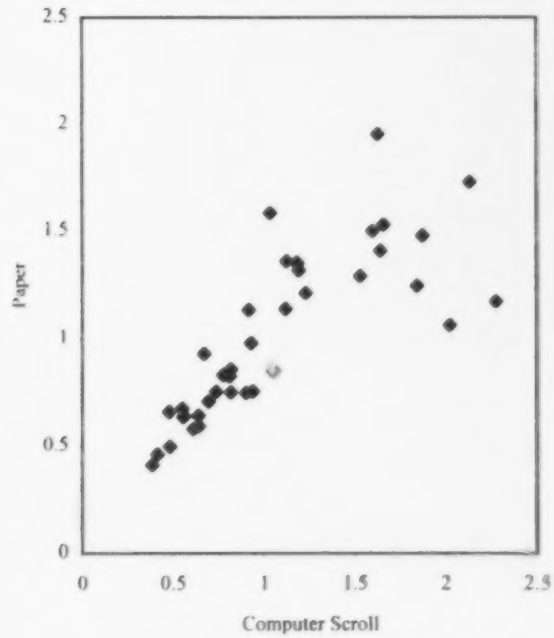
Computer Page Vs. Paper



Computer Page Vs. Re-estimated Computer Page



Computer Scroll Vs. Paper



Computer Scroll Vs. Re-estimated Computer Scroll

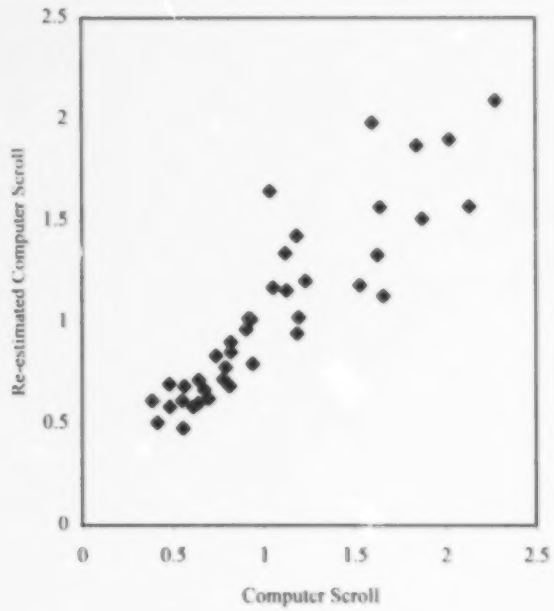
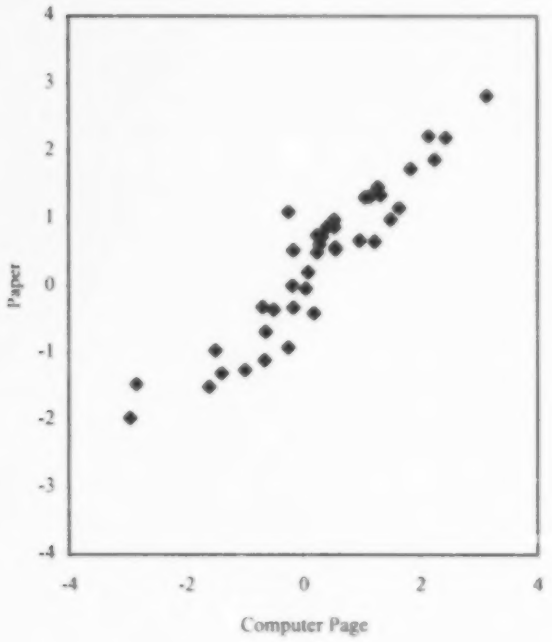
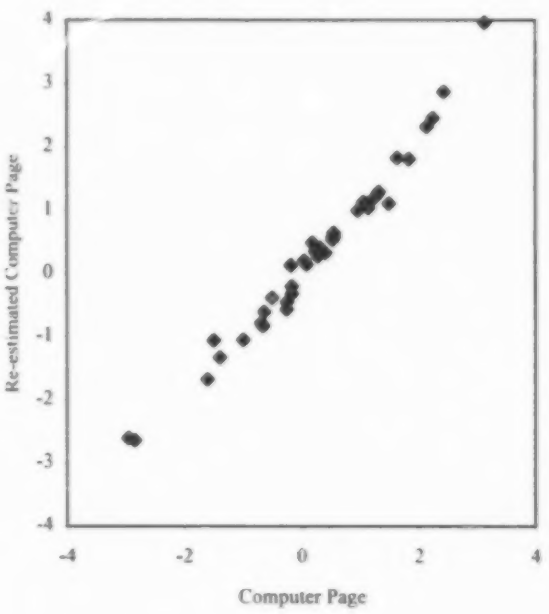


Figure 13. B Parameters for Comparability 2 Science Reasoning
(Computer Vs. Paper and Computer Vs. Re-estimated Computer).

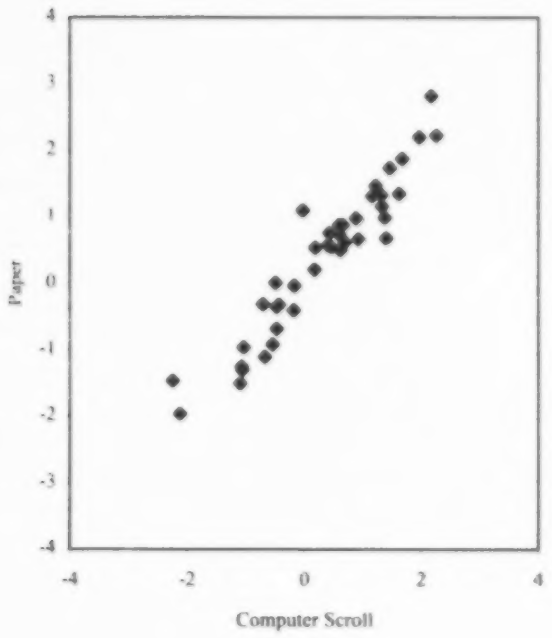
Computer Page Vs. Paper



Computer Page Vs. Re-estimated Computer Page



Computer Scroll Vs. Paper



Computer Scroll Vs. Re-estimated Computer Scroll

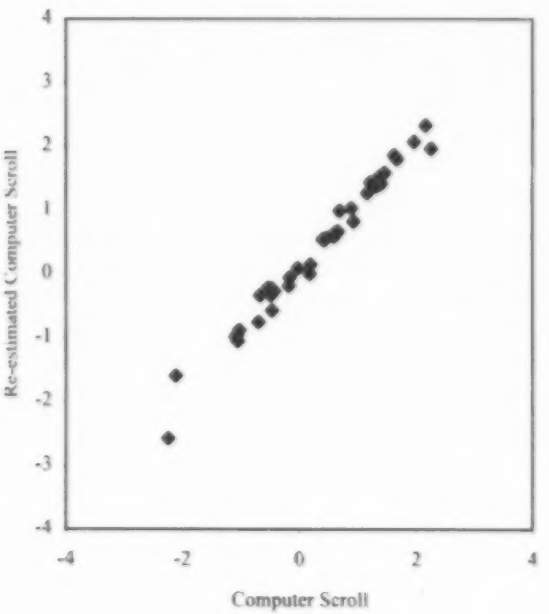
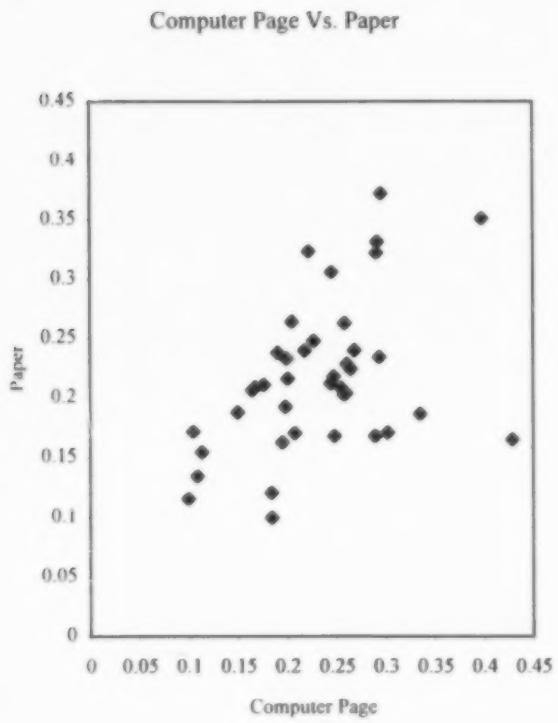
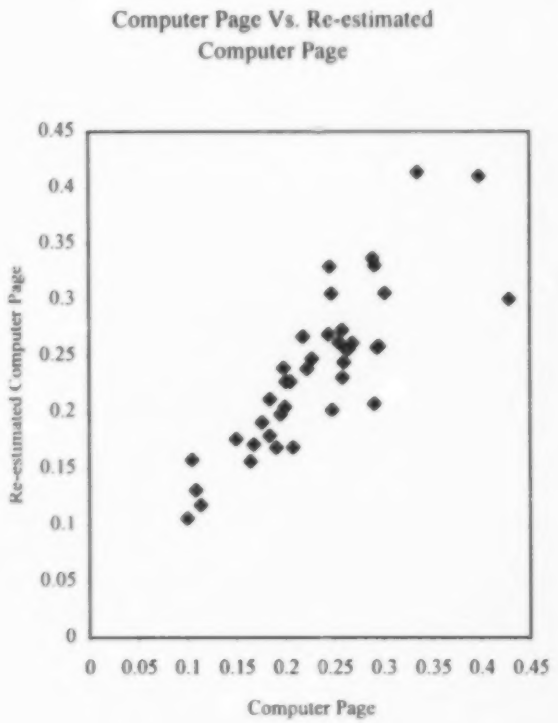
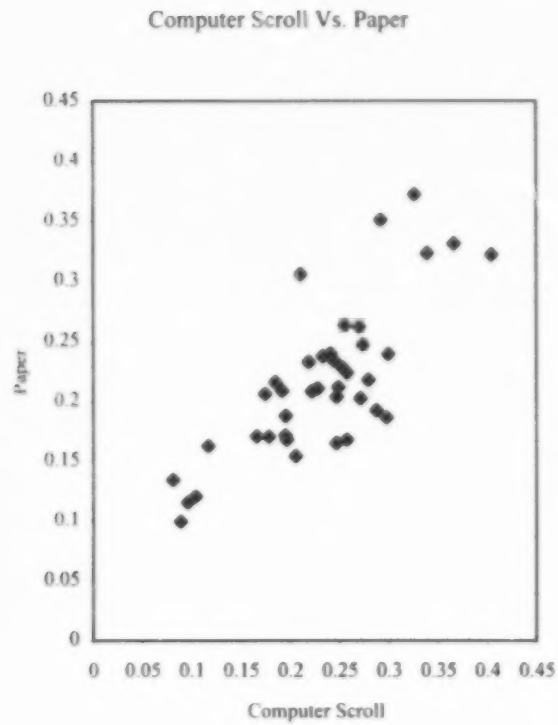


Figure 14. C Parameters for Comparability 2 Science Reasoning
(Computer Vs. Paper and Computer Vs. Re-estimated Computer).







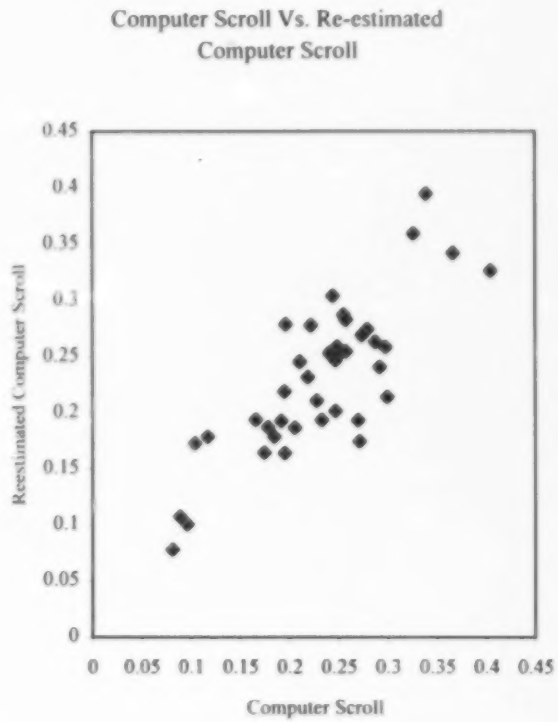

Computer Scroll Vs. Re-estimated Computer Scroll
Reestimated Computer Scroll
Computer Scroll
0 0.05 0.1 0.15 0.2 0.25 0.3 0.35 0.4 0.45

Figure 15. Predicted Test-Retest Reliabilities for Reading Page

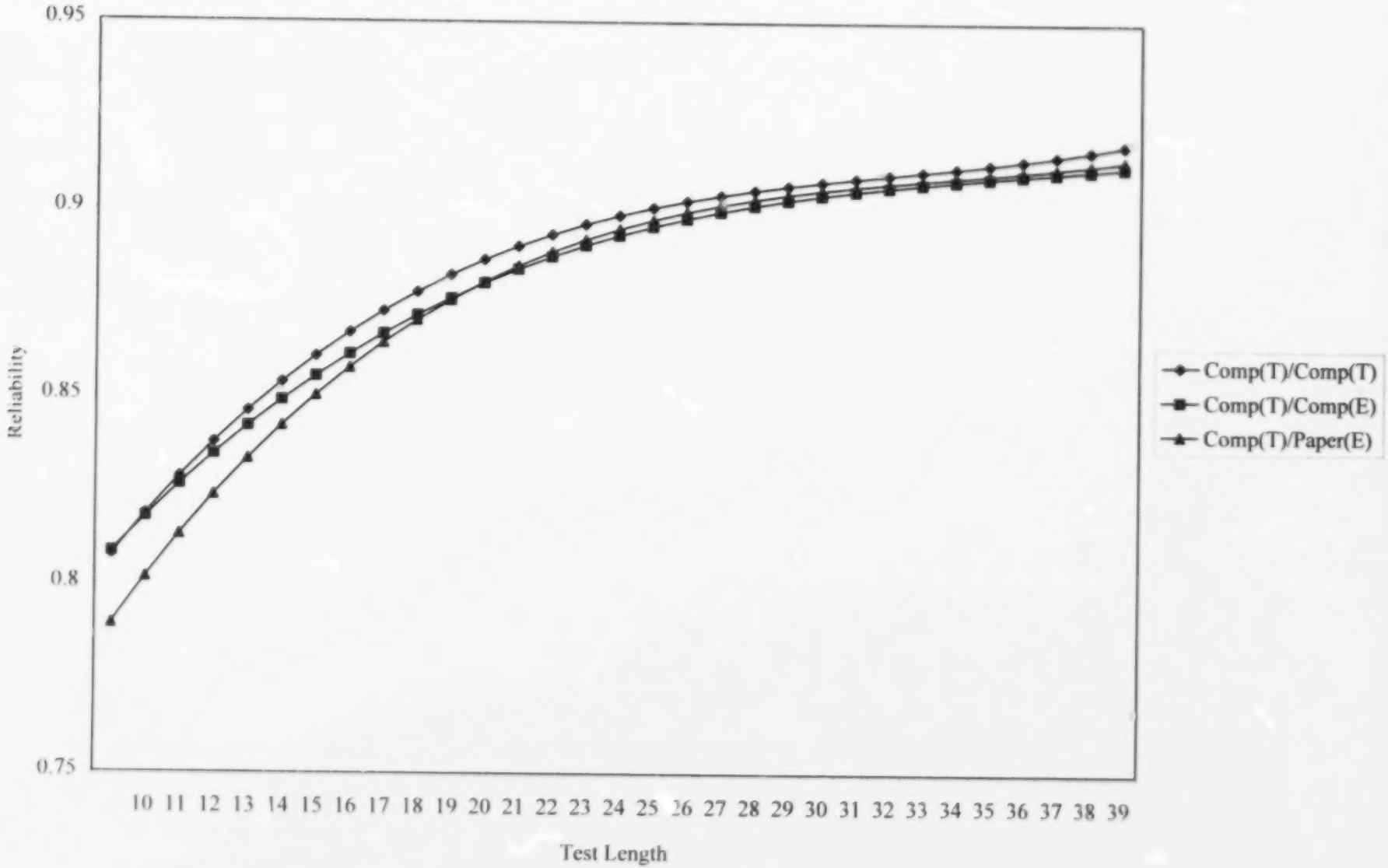


Figure 16. Predicted Test-Retest Reliabilities for Reading Scroll.

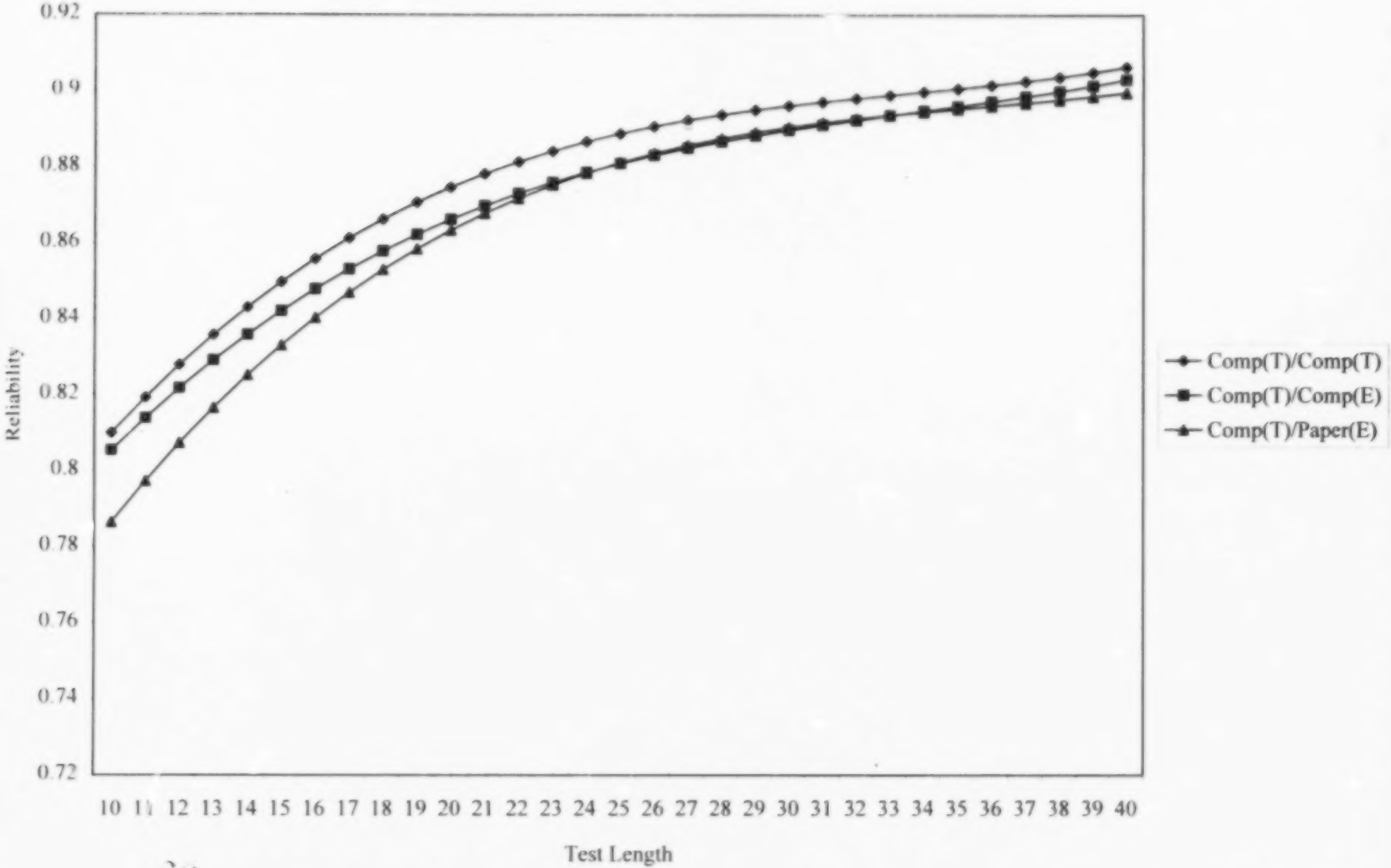


Figure 17. Predicted Test-Retest Reliabilities for Science Page

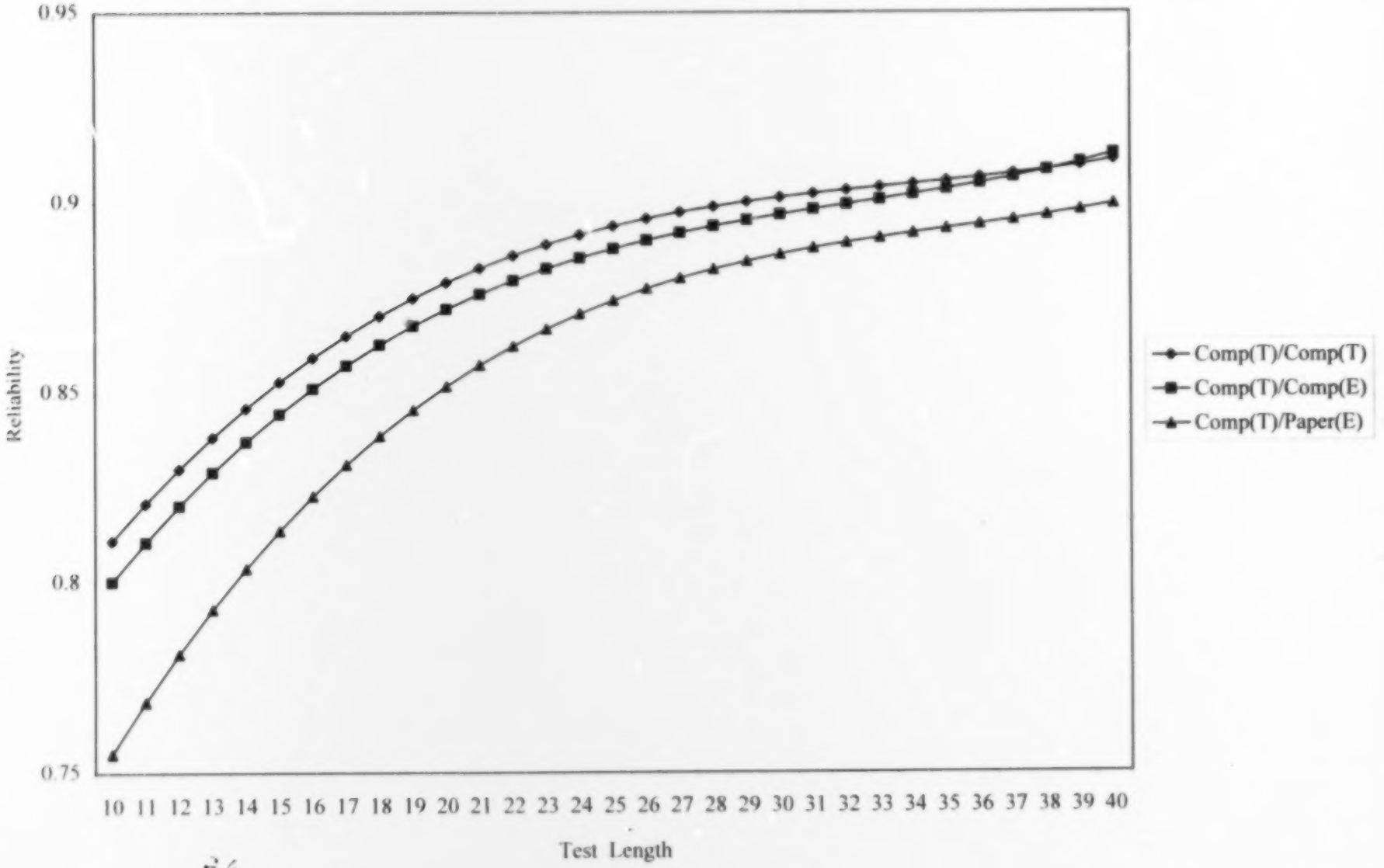


Figure 18. Predicted Test-Retest Reliabilities for Science Scroll